Globale genomische Epidemiologie von *Clostridioides difficile*

Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades einer

Doktorin der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Martinique Frentrup
aus Duisburg

1. Referent: Prof. Dr. Ulrich Nübel
2. Referent: Prof. Dr. Michael Steinert
eingereicht am: 24.06.2020
mündliche Prüfung (Disputation) am: 27.10.2020
Druckjahr 2021

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Frentrup, M., Zhou, Z., Steglich, M., Meier-Kolthoff, JP., Göker, M., Riedel, T., Bunk, B., Spröer, C., Overmann, J., Blaschitz, M., Indra, A., von Müller, L., Kohl, TA., Niemann, S., Seyboldt, C., Klawonn, F., Kumar, N., Lawley, TD., García-Fernández,S., Cantón, R., del Campo, R., Zimmermann, O., Groß, U., Achtmann, M., Nübel, U. 2020. A publicly accessible database for *Clostridioides difficile* genome sequences supports tracing of transmission chains and epidemics. *In revision*

García-Fernández,S., Frentrup, M., Steglich, M., Gonzaga, A., Cobo, M., López-Fresneña, N., Cobo, J., Morosini, MI., Cantón, R., del Campo, R., Nübel, U. 2019. Whole-genome sequencing reveals nosocomial *Clostridioides difficile* transmission and a previously unsuspected epidemic scenario. *Scientific Reports* doi: 10.1038/s41598-019-43464-4

Numberger, D., Riedel, T., McEwen, G., Nübel, U., Frentrup, M., Schober, I., Bunk, B., Spröer, C., Overmann, J., Grossart, HP., Greenwood, AD. 2019. Genomic analysis of three *Clostridioides difficile* isolates from urban water sources. *Anaerobe* doi: 10.1016/j.anaerobe.2019.01.002

Berger, FK., Gfrörer, S., Becker, SL., Baldan, R., Cirillo, DM., Frentrup, M., Steglich, M., Engling, P., Nübel, U., Mellmann, A., Bischoff, M., Gärtner, B., von Müller, L. 2019. Hospital outbreak due to *Clostridium difficile* ribotype 018 (RT018) in Southern Germany. *International Journal of Medical Microbiology* doi: 10.1016/j.ijmm.2019.03.001

Tagungsbeiträge

Frentrup, M., Zhou, Z., Steglich, M., Meier-Kolthoff, JP., Göker, M., Riedel, T., Bunk, B., Spröer, C., Overmann, J., Blaschitz, M., Indra, A., von Müller, L., Kohl, TA., Niemann, S., Seyboldt, C., Klawonn, F., Kumar, N., Lawley, TD., Garcia-Fernandez, S., Canton, R., del Campo, R., Zimmermann, O., Groß, U., Achtmann, M., Nübel, U.: A publicly accessible database for *Clostridioides difficile* genome sequences supports tracing of transmission chains and epidemics. (Vortrag) 6. Gemeinsame Jahrestagung der Deutschen Gesellschaft für Hygiene und Mikrobiologie (DGHM) e. V. zusammen mit der Vereinigung für Allgemeine und Angewandte Mikrobiologie (VAAM) e. V., Leipzig, 2020

Frentrup, M., Sergeant, MJ., Zhou, Z., Alikhan, NF., Blaschitz, M., Indra, A., von Müller, L., Steglich, M., M., Riedel, T., Bunk, B., Spröer, C., Overmann, J., Zimmermann, O., Groß, U., Achtmann, M., Nübel, U.: A Genome Database for *Clostridioides difficile* implemented in EnteroBase. (Vortrag) Jahrestagung der Vereinigung für Allgemeine und Angewandte Mikrobiologie (VAAM) e. V., Wolfsburg, 2018

Frentrup, M., Sergeant, MJ., Zhou, Z., Alikhan, NF., Blaschitz, M., Indra, A., von Müller, L., Steglich, M., M., Riedel, T., Bunk, B., Spröer, C., Overmann, J., Zimmermann, O., Groß, U., Achtmann, M., Nübel, U.: A Genome Database for *Clostridioides difficile* implemented in EnteroBase. (Vortrag) 70. Jahrestagung der Deutschen Gesellschaft für Hygiene und Mikrobiologie (DGHM) e. V., Bochum, 2018

Frentrup, M., Sergeant, MJ., Zhou, Z., Alikhan, NF., Blaschitz, M., Indra, A., von Müller, L., Steglich, M., M., Riedel, T., Bunk, B., Spröer, C., Overmann, J., Zimmermann, O., Groß, U., Achtmann, M., Nübel, U.: A Genome Database for *Clostridioides difficile* implemented in EnteroBase. (Vortrag) 28th European Congress of Clinical Microbiology and Infectious Diseases (ECCMID), Madrid, 2018

Thiel, N., Frentrup, M., Junker, V., Siller, P., Amon, T., Rösler, U., Nübel, U.: *Clostridioides difficile* in poultry manure. (Vortrag) 71. Jahrestagung der Deutschen Gesellschaft für Hygiene und Mikrobiologie (DGHM) e. V., Göttingen, 2019

Inhaltsverzeichnis

Vorveröffentlichungen der Dissertation **I**

Zusammenfassung **V**

Abbildungsverzeichnis **VII**

Tabellenverzeichnis **XIII**

1 Einleitung **1**

1.1 Das Darmbakterium *Clostridioides difficile* . . . 2

1.1.1 Phänotypische und genotypische Eigenschaften . . . 3

1.1.2 Pathogenese . . . 3

1.1.3 Diagnostik und Behandlung . . . 4

1.1.4 Surveillance . . . 4

1.2 Typisierungsmethoden für *C. difficile* Isolate . . . 5

1.2.1 Die PCR-Ribotypisierung . . . 6

1.2.2 Sequenzbasierte Typisierungsmethoden . . . 7

1.3 Molekulare Epidemiologie von *C. difficile* . . . 9

1.3.1 Übersicht über die Epidemiolgie der letzten Jahre . . . 9

1.3.2 Detektion von Transmissionswegen durch genomische Analysen . . . 10

1.4 Die Softwareplattform EnteroBase . . . 11

1.5 Zielsetzung . . . 12

2 Methodik **13**

2.1 Generierung der Sequenzdaten . . . 13

2.1.1 Stammsammlung . . . 13

2.1.2 PCR Ribotypisierung und Literaturrecherche . . . 14

2.1.3 DNA Extraktion und Sequenzierung . . . 15

2.2 Prozessierung der Sequenzdaten . . . 15

2.2.1 Qualitätsprüfung und Auswahl der Sequenzdaten . . . 15

2.2.2 Bioinformatische Analysen . . . 18

2.3 Statistische Untersuchungen . . . 25

2.3.1 Evaluierung des cgMLST Schemas . . . 25

2.3.2 Die Populationsstruktur von *C. difficile* . . . 26

2.3.3 Quantitativer Vergleich der SNP- und cgMLST-Analyse . . . 26

2.3.4 Lokale und globale Epidemiologie . . . 26

3 Ergebnisse **29**

3.1 Evaluierung des cgMLST Schemas in EnteroBase . . . 29

3.2 Die Populationsstruktur von *Clostridioides difficile* . . . 30

3.2.1 Detektion von pandemischen Stämmen . . . 30

3.2.2 Die hierarchische Clusterung in HC150 Cluster - eine Alternative zur PCR Ribotypisierung . . . 31
3.2.3 Beispielanwendung der Typisierung durch EnteroBase . . . 36
3.3 Quantitativer Vergleich der cgMLST- und SNP-Analyse . . . 37
3.4 Anwendung der cgMLST-Analyse auf lokale und globale Epidemiologie . . . 39
3.4.1 Differenzierung eines Rezidivs von einer Neuinfektion durch *C. difficile* . . . 39
3.4.2 Anwendung der hierarchischen Clusterung zur Detektion von lokalen und regionalen Ausbrüchen von *C. difficile* . . . 40
3.4.3 Verbreitung von *C. difficile* durch Ausbringung von tierischem Dung auf landwirtschaftlich genutzte Flächen . . . 42
3.4.4 Analyse der globalen Verbreitung von *C. difficile* . . . 44
3.5 Herausforderungen der genomischen Epidemiologieanalysen . . . 49
3.5.1 Nahezu identische Genome ohne epidemiologischen Zusammenhang . . . 49
3.5.2 SNP-Analyse von hoch diversen Datensätzen . . . 53

4 Diskussion **57**
4.1 Bewertung und Vergleich der bioinformatischen Methoden in der genomischen Epidemiologie 58
4.1.1 Kritische Betrachtung der verfügbaren Methoden . . . 58
4.1.2 Eignung der cgMLST für die molekulare Epidemiologie . . . 61
4.1.3 Aussicht und Entwicklungspotential der Methoden . . . 63
4.2 Kurzzeitige Evolution und örtliche Ausbreitung von *C. difficile* . . . 64
4.2.1 Erfassung der globalen Populationsstruktur auf verschiedenen Ebenen . . . 64
4.2.2 Detektion von nah verwandten Isolaten . . . 68
4.2.3 Limitierungen der Ausbruchsdetektion . . . 71
4.2.4 Aussicht für weitere Analysen der Populationsstruktur von *C. difficile* . . . 73
4.3 Fazit . . . 74

A Isolatelisten **75**

B Höhere Populationsebenen von *C. difficile* **103**

C Ergänzende Tabellen **107**

Literatur **113**

Danksagung **127**

Zusammenfassung

Clostridioides difficile gilt als Haupterreger von nosokomial auftretender, antibiotikaassoziierter Diarrhö. Die globale Ausbreitung des Erregers sowie direkte Transmissionswege wurden mit Hilfe der Ganzgenomsequenzierung intensiv untersucht, konnten aber bis jetzt nur auf einem begrenzten Datensatz durchgeführt werden. Die auf der Softwareplattform EnteroBase (`http://enterobase.warwick.ac.uk`) verfügbare umfangreiche Datenbank von *C. difficile* Genomen, die im Zuge dieser Arbeit mit aufgebaut und kuratiert wurde, beinhaltete zum Zeitpunkt der Analysen 13.515 Genome und wächst durch tägliches Beziehen von veröffentlichten Sequenzdaten aus Repositorien wie NCBI stetig weiter. Durch die implementierten bioinformatischen Werkzeuge ermöglicht EnteroBase Sequenzdaten einheitlich zu prozessieren und diese anhand eines Kerngenom-MLST (cgMLST) Schemas zu typisieren. Zudem werden Einträge anhand ihrer paarweisen Übereinstimmungen in den cgMLST Allelprofilen in hierarchische Cluster eingeteilt.

Die Ergebnisse dieser Arbeit demonstrieren, dass EnteroBase mit der Einteilung der Genome in hierarchische HC150 Cluster eine einheitliche Typisierungsmethode von *C. difficile* Isolaten bietet, die alternativ zur standardmäßig verwendeten PCR Ribotypisierung angewendet werden kann. Des Weiteren fallen die Genome von Isolaten, die in Publikationen Pandemien zugeordnet wurden, in HC10 Cluster. Der erstmals durchgeführte, quantitative Vergleich mit der standardmäßig zur Detektion von Transmissionsketten verwendeten SNP-Analyse zeigte, dass anhand der cgMLST-Allelprofile vergleichbare genomische Unterschiede zwischen Isolaten festgestellt wurden. Eyre et al. stellten fest, dass Isolate mit einer genomischen Distanz von $\leq$2 SNPs mit 95 %iger Wahrscheinlichkeit einer Transmissionskette angehören. Dieser Wert konnte auch auf die Kerngenom-Allelunterschiede angewendet werden. Die hierarchische Clusterung in EnteroBase fasst Isolate mit einer kettenweisen genomischen Distanz von $\leq$2 Kerngenom-Allelunterschieden in ein HC2 Cluster zusammen. Dadurch konnten anhand der HC2 Cluster rezidivierende *C. difficile* Infektionen von Neuinfektionen unterschieden werden. Zudem wurden die HC2 Cluster herangezogen um retrospektiv Transmissionswege in einem Netzwerk von Krankenhäusern aufzudecken, wobei hier eine signifikante Assoziation zwischen den HC2 Clustern und den beprobten Krankenhäusern beziehungsweise Stationen bestand. Durch die Analyse der HC150 und der HC2 Cluster konnte weiterhin gezeigt werden, dass *C. difficile* Isolate in kontaminierten Dung, der auf landwirtschaftliche Flächen ausgebracht wurde, über mehrere Wochen im Boden überleben konnten und über Staubpartikel in der Atmosphäre verbreitet wurden. Für nahe genomische Verwandtschaften zwischen epidemiologisch nicht zusammenhängenden Isolaten schien der Wert einer genomischen Distanz von $\leq$2 allerdings nicht praktikabel. Isolate mit unterschiedlichem Herkunftsland und weit auseinander liegenden Isolationsdaten zeigten trotz naher Verwandtschaft im Kerngenom große Unterschiede in ihrem akzessorischem Gengehalt, sodass die Zugehörigkeit zum gleichen Ausbruchsklon implausibel erschien.

In der vorliegenden Arbeit konnte anhand der *C. difficile* Datenbank in EnteroBase ein zuvor noch nicht erfasster umfangreicher Einblick in die Populationsstruktur des pathogenen Bakteriums gewonnen werden. Der hier erbrachte Beweis, dass anhand der bioinformatischen Werkzeuge in EnteroBase *C. difficile* Isolate typisiert und Transmissionswege aufgedeckt werden können, ermöglicht auch Wissenschaftlern ohne bioinformatischem Hintergrund in Zukunft entsprechende Analysen in einem globalen Kontext auf standardisierte Weise durchzuführen.

Abbildungsverzeichnis

1.1 **Erfassung von genomischen Unterschieden** durch die MLST- und die SNP-Analyse an einem Beispiel von vier Sequenzvergleichen. **(A)** Bei der MLST-Analyse werden die Sequenzen in Abschnitte eingeteilt, so genannte Loci. Jeder individuell erfassten Sequenz an einer Locus-Stelle wird eine individuelle Allelnummer zugeordnet. Dadurch entsteht für jede Sequenz ein individuelles Allelprofil, anhand derer die vier Sequenzen miteinander verglichen werden können. **(B)** Bei der SNP-Analyse werden nach der Alignierung der Illumina-*Reads* gegen eine Referenzsequenz die von der Referenz abweichenden Basen detektiert (hier rot markiert). Bei einem Vergleich der Sequenzen wird jede Punktmutation als Unterschied gewertet. 8

2.1 **Übersicht der durchgeführten Analysen** für die Illumina-*Reads* ausgewählter Datensätze und deren zugehöriges Ergebnis. Die eingefärbten Analysewege werden in den Abbildungen 2.2, 2.3 und 2.4 detaillierter dargestellt. 19

2.2 **Ablauf der cgMLST-Analyse** der qualitätskontrollierten Illumina-*Reads*. Nachdem die *Reads* auf die EnteroBase-Plattform hochgeladen wurden, wurden diese automatisch duch eine Pipeline assembliert und anschließend die Qualität der resultierenden Assemblierungen überprüft. Für Assemblierungen mit ausreichender Qualität wurden folgend Allelprofile für alle in EnteroBase verfügbaren MLST Schemata bestimmt. Auf Basis der durch die cgMLST Allelprofile bestimmten Kerngenom-Sequenztypen wurden den Assemblierungen hierarchische Cluster auf insgesamt 13 Ebenen zugeordnet (HC0, 2, 5, 10, 20, 50, 100, 150, 200, 500, 950, 2000 und HC2500). Des Weiteren wurden die cgMLST Allelprofile zur Berechnung der paarweisen Distanzen in die Software Bionumerics® importiert. Die berechnete Distanzmatrix wurde dann in R in paarweise Vergleiche der Kerngenom-Allelunterschiede prozessiert. 20

2.3 **SNP-EToKi-Analyse** der qualitätskontrollierten Illumina-*Reads*. Nachdem die *Reads* auf die EnteroBase Plattform hochgeladen und automatisch assembliert wurden, folgt eine Qualitätsüberprüfung der Assemblierungen. Die *Contigs* der Assemblierungen mit ausreichender Qualität wurden gegen eine Referenzsequenz aligniert. Mit dem entstehendem Alignment wurde ein Maximum-Likelihood phylogenetischer Baum berechnet. Anschließend wurden mit Hilfe des Alignments und des Baumes die Rekombinationen bestimmt. Im Folgendem wurden die SNPs, die als Folge einer Rekombination entstanden, in dem Alignment maskiert. Auf Basis dieses Alignments wurden eine SNP-Distanzmatrix berechnet. Hierbei wurden nicht bestimmte Basen sowie die maskierten Rekombinationen paarweise ausgeschlossen. Die entstandene SNP-Distanzmatrix wude in R weiter prozessiert. 22

2.4 **Ablauf der SNP-Analyse** der qualitätskontrollierten Illumina-*Reads*. Nachdem die *Reads* gegen eine Referenzsequenz aligniert wurden, wurde die Konsensussequenz ermittelt. Anschließend wurden alle zu analysierenden Konsensussequenzen mit der für den *Mapping*-Prozess verwendeten Referenzsequenz in ein Alignment zusammen gefasst und ein Maximum-Likelihood phylogenetischer Baum berechnet. Mithilfe des berechneten Baumes und des Alignments wurden folgend die Koordinaten von Rekombinationen ermittelt. Die bestimmten Koordinaten wurden spaltenweise aus dem Alignment entfernt. Anschließend wurde das Alignment auf die Mutationen enthaltenen Stellen gekürzt und die paarweisen SNPs in R berechnet. 23

2.5 **Vereinfachte Darstellung** der Erstellung eines Minimum-Spanning Baumes. Zunächst wird ein gerichteter Arboreszenz Baum erstellt, in dem alle Knoten durch einen Pfad von der Wurzel aus erreichbar sind und die asymmetrischen Zweige die genomische Distanz basierend auf Kerngenom-Allelunterschieden wiedergeben **(A)**. Es werden die Zweige ausgewählt, durch die alle Knoten mit der geringsten Endsumme (Addition aller genomischen Distanzen der ausgewählten Zweige) verknüpfbar sind. Anschließend erfolgt eine lokale Neuzuordnung, um störende Verzweigungen zu entfernen und die Astlängen proportional zur Genomdistanz darzustellen **(B)**. 25

3.1 **Evaluierung des cgMLST Schemas in EnteroBase. (A)** Verteilung der Anzahl an unbestimmten Allelen pro Genom. Der grüne Balken fasst die Ausreißer zusammen. **(B)** Verteilung der Anzahl an unterschiedlichen Allelsequenzen pro Locus. **(C)** Verteilung der Sequenzlängen pro Locus. **(D)** Einfluss der Sequenzlänge auf die Alleldiversität pro Locus. . . 29

3.2 **Phylogenetische Bäume von drei *C. difficile* Pandemien.** Die Bäume wurden auf Grundlage der cgMLST Allelprofile und unter Anwendung des in EnteroBase verfügbaren Rapid-Neighbour-Joining Algorithmus berechnet. Die Farben spiegeln das jeweilige HC10 Cluster wider. CC: cgST Complex; RT: PCR Ribotyp. 31

3.3 **Häufigkeitsverteilung der Kerngenom-Allelunterschiede** zwischen den verfügbaren 13.515 Genomen in EnteroBase. Die Linien geben die Grenzen der hierarchischen Cluster HC150, HC950, HC2000 und HC2500 an den entsprechenden Kerngenom-Allelunterschieden an. 32

3.4 **Vergleich der HC150 Cluster und PCR Ribotypen.** Rapid-Neighbor-Joining phylogenetische Bäume basierend auf cgMLST Allelprofilen von 2.263 Genomen, für die die Information über den PCR Ribotypen zur Verfügung stand. **(A)** Farben zeigen PCR Ribotypen an. **(B)** Farben zeigen CCs an. CC: cgST Complex. 33

3.5 **Phylogenetischer Baum aller 13.515 *C. difficile* Genome** in EnteroBase, basierend auf deren cgMLST Allelprofilen. Zur Berechnung des Baumes wurde der Rapid-Neighbour-Joining Algorithmus verwendet. Die Farben und Zahlen stehen für die entsprechenden CCs, die mindestens zehn Einträge umfassen. Zugehörige PCR Ribotypen werden in den Klammern erwähnt. CC: cgST Complex; RT: PCR Ribotyp. 34

3.6 **Rarefaction Analyse der hierarchischen Cluster** HC150, HC950, HC2000 und HC2500. Die gestrichelten Linien stellen die Extrapolation und damit die Schätzung der vorhandenen hierarchischen Cluster über die schon in EnteroBase erfassten dar. 36

3.7 **Korrelationsanalyse zwischen den genomischen Distanzen der cgMLST- und SNP-Analyse.** Die oberen Graphen basieren auf den cgMLST Allelprofilen aus EnteroBase, die unteren auf denen aus SeqSphere$^+$. Die Größe der Punkte spiegelt die Anzahl der Datenpunkte wider. Die Graphen zeigen die Gegenüberstellung der paarweisen Kerngenom-Allelunterschiede und SNP Distanzen für die folgenden Datensätze: **(A)** Isolate von vier rezidivierenden CDI Patienten, welche an zwei Zeitpunkten beprobt wurden (Anzahl der Isolate, n = 176). **(B)** Isolate von vier kürzlich publizierten Ausbrüchen, darunter ein Ausbruch in einem chinesischem Krankenhaus, der sich über zwei Jahre zog (n = 12) [112] und ein Ausbruch in Süddeutschland, der zwei Krankenhäuser betraf (n = 9) [120]. Zudem beinhaltet der Datensatz noch zwei Ausbrüche aus einem Krankenhaus in Madrid (Spanien), welche die PCR Ribotypen 027 (n = 22) und 106/500 (n = 20) betreffen [55]. **(C)** Isolate, die zwischen 2007 und 2011 in vier Krankenhäusern in Oxfordshire, Vereinigtes Königreich, von CDI erkrankten Patienten isoliert wurden (n = 1.158) [72]. Die genomischen Distanzen werden hier bis zu einem Wert von 10 gezeigt. 37

3.8 **Binäres logistisches Regressionsmodell** angewandt auf den Oxfordshire Datensatz von 1.158 Genomen. Um die Wahrscheinlichkeit zu berechnen, dass zwei Genome sich bei einem gewissen Kerngenom-Allelunterschied in $\leq$2 SNPs unterscheiden, wurden die SNP Distanzen binär codiert (1 wenn $\leq$2 SNPs, 0 wenn >2 SNPs). Die Allelunterschiede wurden als Vorhersagevariabel genutzt. 38

3.9 **Wiederkehrende C. difficile Infektionen.** Minimum-Spanning Bäume basierend auf den cgMLST Allelprofilen der Genome von Isolaten, die aus vier Patienten mit wiederkehrender CDI zu zwei Zeitpunkten isoliert wurden. Die Zahlen geben die Anzahl der cgMLST Allelunterschiede an. Rot: Erster Aufenthalt; blau: Zweiter Aufenthalt. 39

3.10 **Phylogenetische Bäume von vier publizierten Ausbrüchen**, basierend auf Kerngenom-Allelunterschieden [55], [112], [120]. Die Bäume wurden mit dem Rapid-Neighbour-Joining Algorithmus berechnet. Die Farben indizieren die HC2 Cluster. Die Skala, die für einen Kerngenom-Allelunterschied steht, gilt für alle Bäume. CC: cgST Complex; RT: PCR Ribotyp. 40

3.11 **Nachgewiesene Transmissionswege in einem Netzwerk von Krankenhäusern.** Die Farben geben die Stationen wider, 'X' steht für den Zeitpunkt der Diagnose der CDI und die Pfeile indizieren die angenommenen Transmissionswege. Rechts befinden sich die zugehörigen Minimum-Spanning Bäume, welche die genomischen Distanzen zwischen den Isolaten zeigen. Oberes Feld: Patient P1 wurde in Krankenhaus K2 mit CDI diagnostiziert und fünfzehn Tage später in Krankenhaus K3 verlegt. Fünf Tage nach der Verlegung wurde in Krankenhaus K3 Patient P3 mit einem nah verwandten Stamm infiziert, ebenso Patient P2 in Krankenhaus K2 nach sechs Tagen. Beide Patienten lagen im jeweiligen Krankenhaus auf der gleichen Station wie der Erstpatient, welcher wahrscheinlich die Infektionsquelle darstellt. Allerdings gab es bei den Patienten P1 und P2 keine zeitliche Überschneidung, sodass die Übertragung wahrscheinlich über die Umgebung stattfand. Unteres Feld: Die Patienten P4 und P5 wurden am selben Tag mit CDI diagnostiziert, nachdem sie sieben Tage lang gemeinsam hospitalisiert waren, allerdings auf verschiedenen Stationen. Bei einem dritten Patient P6 wurde eine CDI mit einem genomisch identischem Stamm wenige Zeit später diagnostiziert. Obwohl die Diagnose in einem anderen Krankenhaus stattfand, liegt eine mögliche Transmission vor, da Patient P6 vorher im gleichen Krankenhaus wie die Patienten P4 und P5 hospitalisiert war. . 41

3.12 **Verteilung der Isolate** in den Proben des SOARiAL Projekts (n = 191). Die Farben geben die den Genomen in EnteroBase zugeordneten CCs an. Die Wochenangaben der gedüngten Bodenproben geben die Zeit nach Ausbringung des Düngers auf den Boden an. 43

3.13 **Minimum-Spanning Bäume** der HC2 Cluster 1206 und 1232, die sowohl Genome von Isolaten aus dem SOARiAL Projekt als auch Genome von klinischen Isolaten in EnteroBase umfassten. Die Farben indizieren die Isolationsquelle des zugehörigen Isolats. 43

3.14 **Kleinster Kerngenom-Allelunterschied** für jedes Genom in EnteroBase mit Länderinformation zu einem anderen Genom aus dem gleichen Land („inländisch“) und zu einem Genom aus einem anderen Land („länderübergreifend“), dargestellt bis zu einer Distanz von <200 (n = 10.957). Die Farbe gibt an, ob dieser Wert ≤2 (rot) oder >2 (blau) Kerngenom-Allelunterschiede beträgt. Die Boxplots geben Information über Median, unteres und oberes Quartil. 45

3.15 **Boxplots der paarweisen Jaccard Distanzen im akzessorischem Genom und der Zeitspanne zwischen den Isolaten** für paarweise Vergleiche mit ≤2 Kerngenom-Alllelunterschieden (Anzahl angegeben durch „n“). Vorauswahl: 1.004 Isolate, die eine nahe genomische Verwandtschaft zu einem Isolat aus einem anderen Land zeigten; Stichproben: 1.000 zufällig ausgewählte Einträge in EnteroBase für die die paarweisen Kerngenom-Alleleunterschiede bestimmt wurden; Ausbrüche: Vier publizierte Ausbrüche; Regional: Oxfordshire Datensatz aus der Publikation von Eyre et al. [72]; Knetsch 2018 und Knight 2019: Isolate, denen in der Publikation eine nahe genomische Verwandtschaft zu einem Isolat aus einem anderen Land nachgewiesen wurde [56], [87]. **(A)** Die An-/Abwesenheitsunterschiede im akzessorischem Genom wurden basierend auf den wgMLST Allelprofilen abzüglich der cgMLST Loci der Genome aus EnteroBase bestimmt. **(B)** Die An-/Abwesenheitsunterschiede im akzessorischem Genom wurden durch die *k-mer* basierte PopPUNK Analyse der Assemblierungen berechnet. **(C)** Für paarweise Vergleiche mit Jahresinformation für beide Isolate wurde die Zeitspanne zwischen den Probennahmen berechnet. 46

3.16 **Darstellung von Mapping-Artefakten** in den Illumina-*Reads* der Isolate von HC2 Cluster 1. Dieses Cluster umfasst sieben Einträge in EnteroBase. Die Illumina-*Reads* dieser sieben Isolate wurden gegen die Referenzsequenz R20291 gemappt. Die linken Abbildungen zeigen die Abdeckungen der Sequenzen bei einer Durchführung des *Read-Mappings* mit Standardeinstellungen (*seed*-Länge = 19), die rechten entsprechend bei einer gewählten *seed*-Länge von 30. **(A)** Nahaufnahme der Abdeckung der Illumina-*Reads* an zwei Abschnitten der Referenzsequenz, an denen SNPs zwischen den Sequenzen detektiert wurden. Die Höhe der grauen Blöcke, die das Nukleotid an dieser Stelle repräsentieren, spiegelt die Abdeckung der Illumina-*Reads* an dieser Stelle wider. Eingefärbte Blöcke zeigen einen Unterschied in der jeweiligen Sequenz zur Referenzsequenz an, wobei die Farbe der Blöcke der an dieser Stelle befindlichen DNA-Base entspricht (Adenin=grün; Guanin=gelb; Cytosin=blau; Thymin=rot). Rot markiert sind die Sequenzen, bei denen eine auffällig hohe Abdeckung detektiert wurde. Dies ist an der Skala links neben den grauen Balken zu sehen. **(B)** Darstellung der kompletten Tiefe der Illumina-*Reads* des Isolats CD-16-00185, die an den in Abbildung **A.** dargestellten Abschnitten aligniert wurden. Jeder horizontale Strich stellt einen Illumina-*Read* dar. Die rote Markierung zeigt die Bereiche des Genoms, in denen in Abbildung **A.** die auffällig hohe Abdeckung detektiert wurde. Die Farbe der vertikalen Linien entspricht der in den Illumina-*Reads* an dieser Position befindlichen Base. Die Abbildung wurde mithilfe der *IGV-Web* Anwendung erstellt (`https://igv.org/app/`). 51

B.1 **Phylogenetischer Baum aller 13.515 *C. difficile* Genome** in EnteroBase, eingefärbt nach HC950 Cluster. Der Baum basiert auf den cgMLST Allelprofilen der Genome und wurde mit dem Rapid-Neighbour-Joining Algorithmus berechnet. 103

B.2 **Phylogenetischer Baum aller 13.515 *C. difficile* Genome** in EnteroBase, eingefärbt nach HC2000 Cluster. Der Baum basiert auf den cgMLST Allelprofilen der Genome und wurde mit dem Rapid-Neighbour-Joining Algorithmus berechnet. 104
B.3 **Phylogenetischer Baum aller 13.515 *C. difficile* Genome** in EnteroBase, eingefärbt nach HC2500 Cluster. Der Baum basiert auf den cgMLST Allelprofilen der Genome und wurde mit dem Rapid-Neighbour-Joining Algorithmus berechnet. 105

Tabellenverzeichnis

2.1 **Qualitätsparameter** für die *C. difficile* Assemblierungen in EnteroBase. Mbp: Megabasenpaare; Kbp: Kilobasenpaare; N50: Anzahl an *Contigs* mit einer Länge, die über 50 % der gesamten Genomsequenz liegt; N: nicht eindeutig bestimmbare Base 15

2.2 **Anzahl der Isolate** isoliert aus vier Patienten mit wiederkehrender CDI. Die Wochenangabe bezieht sich auf die Zeitspanne zwischen der Diagnose während des ersten und zweiten Aufenthaltes des Patienten im Krankenhaus. 16

2.3 **Anzahl der Isolate** von Patienten mit CDI, die in einem der sechs beprobten Krankenhäuser hospitalisiert waren. Die Isolate wurden über einen Zeitraum von drei Perioden gesammelt. . 16

2.4 **Anzahl der untersuchten Isolate** aus den verschiedenen Proben des SOARiAL Projekts. Bis auf das Staubisolat wurden die Isolate aus Proben des Feldversuches isoliert. Das Staubisolat wurde aus einer Probe des Windkanalversuchs isoliert. 17

2.5 **Referenzsequenzen** die für die Alignierung der Illumina-*Reads* in der SNP-Analyse beziehungsweise der *Contigs* der Assemblierungen in der SNP-EToKi-Analyse verwendet wurden. 23

2.6 **Kontingenztafel** für die Genome der Isolate, die in einem regionalen Netzwerk von Krankenhäusern isoliert wurden. 133 Isolate dieser Studie wurden durch die cgMLST-Analyse in 23 HC2 Cluster eingeteilt. Die Tabelle zeigt die Verteilung der Isolate pro HC2 Cluster über die beprobten Krankenhäuser an. 27

3.1 **Statistiken der Evaluierung des cgMLST Schemas.** Die Zahlen beziehen sich auf die cgMLST Profile der 13.515 Genome, die zu dem Zeitpunkt der Analyse in EnteroBase zur Verfügung standen. Die Alleldiversität beschreibt die Anzahl der verschiedenen Allelsequenzen pro Locus. Das hierarchische Cluster HC0 fasste Einträge zusammen, dessen cgMLST Profile sich in keinem der bestimmten Allele unterschieden. 30

3.2 **Übersicht über CCs** mit ≥ 100 Einträgen. Die Information über Länder, Isolationsjahr und Quelle der Isolate wurde aus den in EnteroBase verfügbaren Metadaten gewonnen. 35

3.3 **Analyse von drei *C. difficile* Wasserisolaten.** Die Information über die nächsten Verwandten wurden durch die verfügbaren Metadaten in EnteroBase bestimmt. CC: cgST Complex; NA: nicht verfügbar. 36

3.4 **Übersicht der HC2 Cluster in den größten CCs.** Die Bezeichnung HC2 global steht für HC2 Cluster, die mindestens aus drei Genomen von Isolaten bestehen, welche mindestens aus zwei Ländern stammen. 44

3.5 **Anzahl an paarweisen Vergleichen $\leq$2 Kerngenom-Alleluntерschieden mit identischen akzessorischen Genomen** bezogen auf den Gengehalt. Vorauswahl: 1.004 Isolate, die eine nahe genomische Verwandtschaft zu einem Isolat aus einem anderen Land zeigten; Stichproben: 1.000 zufällig ausgewählte Einträge aus der EnteroBase-Datenbank; Ausbrüche: Vier publizierte Ausbrüche; Regional: Oxfordshire Datensatz; Knetsch 2018 und Knight 2019: Isolate, denen in der Publikation eine nahe genomische Verwandtschaft zu einem Isolat aus einem anderen Land nachgewiesen wurde [56], [87]. 47

3.6 **Kerngenom und SNP Distanzen zwischen Genomen des HC2 Cluster 1.** 50
3.7 **SNP Distanzen** für HC2 Cluster 1 und 76, basierend auf *Read-Mappings* gegen verschiedene Referenzsequenzen. Eine Tabelle mit den Ergebnissen für alle 15 untersuchten HC2 Cluster befindet sich in Anhang B 52
3.8 **Anzahl der Isolate in paarweisen Vergleichen** mit einer genomischen Distanz von 0, 1, 2 und *leq*2, ermittelt durch die cgMLST- und die EToKi-SNP Analyse für den ausgewählten Datensatz von 1004 Isolaten aus Kapitel 3.4.4. 53
3.9 **Anzahl der paarweisen Vergleiche und der involvierten Isolate** mit einer genomischen Distanz von $\leq$2. Die Distanzen wurden neben der cgMLST- und SNP-EToKi-Analyse mit verschiedenen Variationen der SNP-Analyse berechnet. Der analysierte Datensatz bestand aus 816 Genomen welche eine nahe Verwandtschaft von $\leq$2 Kerngenom-Allelunterschieden zu einem anderen Genom in EnteroBase, dessen zugehöriges Isolat in einem anderen Land isoliert wurde, aufzeigte (Kapitel 3.4.4; eigentlich 1.004 Genome, aber nur für 816 konnten die Illumina*Reads* heruntergeladen werden). ClonalFrameML und RecHMM wurden zur Detektion der Rekombinationen verwendet. 54

A.1 **Isolateliste des Datensatzes von Patienten mit rezidivierender CDI** (Kapitel 2.2.1). Die zugehörigen Sequenzen wurden im Europäischen Nukleotid Archiv unter der *Study Accession number* PRJEB33768 hinterlegt. 75
A.2 **Isolateliste des Datensatzes aus einem Netzwerk von Krankenhäusern** (Kapitel 2.2.1). Die Isolate wurden aus humanen Proben aus Deutschland isoliert und die zugehörige Krankenhaus- bzw. Stationsinformation wurde anonymisiert. Die zugehörigen Sequenzen wurden im Europäischen Nukleotid Archiv unter der*Study Accession number* PRJEB33779 hinterlegt. 80
A.4 **Liste der Isolate mit PCR Ribotypen Information** (Kapitel 2.2.1). Die zugehörigen Sequenzen wurden jeweils im Europäischen Nukleotid Archiv unter der *Study Accession number* hinterlegt. 90
A.5 **Liste der Isolate, die im Rahmend es SOARiAL Projektes** isoliert wurden (Kapitel 2.2.1). Das Staubisolat wurde aus einer Probe des Windkanalversuchs isoliert, die anderen aus den entsprechenden Proben des Feldversuchs. 98

C.1 **SNP Distanzen** basierend auf *Read-Mappings* gegen verschiedene Referenzsequenzen für paarweise Vergleiche zwischen Genomen der Krankenhausisolate aus Kapitel 3.4.2 und einem anderen Genom in EnteroBase. Diese Beziehungen konnten in insgesamt 15 HC2 Cluster gefunden werden. 107
C.2 **Kontingenztafel** für die Genome der Isolate, die in einem regionalen Netzwerk von Krankenhäusern isoliert wurden. 133 Isolate dieser Studie wurden durch die cgMLST-Analyse in 23 HC2 Cluster eingeteilt. Die Tabelle zeigt die Verteilung der Isolate pro HC2 Cluster über die Stationen der beprobten Krankenhäuser an. 110

Kapitel 1

Einleitung

In den finalen Wochen dieser Arbeit wurde die Gesellschaft weltweit durch das sich schnell verbreitende infektiöse Coronavirus (SARS-CoV-2) mit einer herausfordernden Situation konfrontiert. Die drastischen und bis zu diesem Zeitpunkt in diesem Ausmaße noch nie durchgeführten Maßnahmen wirkten sich sowohl auf die Wirtschaft als auch auf das soziale, alltägliche Leben aus[1], [2]. Dadurch, dass die gesamte Bevölkerung von den Maßnahmen betroffen ist, stieg seit Beginn der Pandemie das Interesse an infektiösen Krankheiten und deren Epidemiologie. Zudem fordert die Gesellschaft eine schnelle und eindeutige Interpretation der Falldaten, sodass getroffene Maßnahmen umgehend angepasst werden können.
Besonders in der aktuellen Situation zeigt sich die Genomik als hilfreiche Methode, um Genome des Coronavirus weltweit miteinander vergleichen und dadurch mögliche Transmissionswege nachverfolgen zu können[3]. Wissenschaftler wollen dadurch eine weitere beziehungsweise erneute Ausbreitung des Coronavirus vermeiden. Seitdem das erste SARS-CoV-2 Genom im Januar 2020 sequenziert wurde, stieg die Zahl schnell auf über 32.000 sequenzierte Genome (Stand: Ende Mai 2020[3]). Innerhalb eines Landes konnten Forscher anhand von Genomvergleichen schon erfolgreich den zeitlichen und örtlichen Ursprung des Pandemieausbruchs in ihrem Land ermitteln[4]. Durch den Vergleich von genomischen Daten kann schnell festgestellt werden, welche Route einer Transmission plausibel ist (geringe Unterschiede) und welche nicht (starke Unterschiede). Besonders bei sich international ausbreitenden Infektionskrankheiten ist der Austausch beziehungsweise der Vergleich der genomischen Daten zwischen den Ländern bedeutend, um weitere Ausbreitungen in andere Länder zu vermeiden. Dabei ist die detaillierte Dokumentation der Patientendaten und deren Kontaktpersonen unumgänglich. Das Auftreten von asymptomatischen Fällen ist sowohl während der Coronavirus-Pandemie, als auch bei anderen infektiösen Krankheiten besonders herausfordernd, da diese meist nicht getestet und dadurch Genome nicht erfasst werden, sodass es zu Lücken in den nachzubildenden Transmissionsketten kommen kann[3].
Obwohl die länderübergreifende Ausbreitung von infektiösen Krankheiten die Epidemiologen schon seit Jahren vor eine große Herausforderung stellt, werden genomische Analysen nicht standardisiert durchgeführt und der internationale Austausch der Daten gestaltet sich als problematisch. Das Angebot an bioinformatischen Werkzeugen ist vielfältig und erschwert so eine einheitliche Untersuchung der Sequenzdaten. Zudem benötigt es meist bioinformatische Expertise um die Werkzeuge anwenden zu können. Aufgrund dessen ist die Nachfrage nach einer Option, die den Vergleich genomischer Daten auf eine standardisierte Art auch für Nicht-Bioinformatiker im internationalem Rahmen ermöglicht, groß. Durch eine zentrale Plattform wie EnteroBase, die neben standardisierten bioinformatischen Werkzeugen auch umfangreiche Datenbanken für ausgewählte pathogene Bakterien beinhaltet, könnte die Kommunikation zwischen Epidemiologen und somit auch eine frühzeitige Erkennung von Pandemien und Epidemien verbessert werden.

Patienten, die an COVID-19 erkranken, werden oft mit Antibiotika behandelt, um mögliche folgende bakterielle Infektionen vorzubeugen[5], [6]. Virusinfektionen der Atemwege führen häufig zu bakteriellen

Lungenentzündungen, sodass COVID-19 Patienten teilweise nicht direkt an der Virusinfektion an sich, sondern an der nachfolgenden bakteriellen Infektion versterben[7]. Zudem können durch die übermäßige Verabreichung von Antibiotika während einer Coronavirus-Infektion pathogene Bakterien, die eine Resistenz gegen das verabreichte Mittel besitzen, eine sekundäre Infektion auslösen[5], [8]. Antibiotikaresistente Bakterien verursachen allein in den USA jährlich 2.8 Millionen Infektionen, die in mehr als 35.000 Fällen zum Tode führen[9] und stellen durch ihr ubiquitäres Vorkommen ein stetiges Risiko für länderübergreifende Pandemien dar[10]–[12].
Antibiotika, die COVID-19 Patienten verabreicht werden, zeigen eine starke Assoziation zu *Clostridioides difficile* Infektionen (CDI)[13] und erste Fälle einer Co-Infektion durch das Coronavirus und *C. difficile* wurden bereits bestätigt[8]. *C. difficile* gilt als Hauptursache nosokomialer, antibiotikassoziierter Diarrhö und steht in Deutschland an vierter Stelle der am häufigsten auftretenden Infektionskrankheiten[14]. Neben immer wieder lokal auftretenden kleineren Ausbrüchen demonstrierte das internationale Aufkommen des hypervirulenten PCR Ribotypen 027 die Fähigkeit von *C. difficile* sich auch über Landesgrenzen hinaus auszubreiten[15], [16]. Aufgrund dessen besteht sowohl in der Forschung als auch in Gesundheitseinrichtungen wie Krankenhäusern ein großes Interesse an einer Möglichkeit *C. difficile* Ausbrüche effektiv zu erfassen und einen umfangreichen Einblick in die internationale Populationsstruktur des pathogenen Bakteriums zu bekommen.

1.1 Das Darmbakterium *Clostridioides difficile*

Das Bakterium *Clostridioides difficile*, früher *Clostridium difficile*, wurde erstmals im Jahre 1935 von Hall und O'Toole unter dem Namen *Bacillus difficilis* beschrieben[17]. Schon damals wurde mit dem Namensteil *'difficilis'* die komplizierten Kultivierungsbedingungen des anaeroben Bakteriums erfasst, wobei es im Laufe der Zeit aufgrund der Stäbchenbildung und der grampositiven Färbung zunächst dem Genus *Clostridium* zugeordnet wurde[18]. Aufgrund von phylogenetischen und phänotypischen Unterschieden zu den anderen Spezies des *Clostridium*-Genus, wurde 2016 ein neuer Genus namens *Clostridioides* unter der Familie Clostridiaceae vorgeschlagen[19]. Neben der jetzt als *Clostridioides difficile* bekannten Spezies wurde außerdem das nächste verwandte Bakterium, *Clostridium mangenotii*, dem neuen Genus zugeordnet und wurde fortan *Clostridioides mangenotii* genannt[19].
Anfangs wurde *C. difficile* als harmloses Darmbakterium wahr genommen. Erst mit der Anwendung von Clindamycin als Breitband-Antibiotikum und der dadurch aufkommenden pseudomembranösen Kolitis wurde *C. difficile* als Ursache von antibiotikaassoziierter Diarrhö erkannt[20], [21]. Folgende Studien zeigten weiterhin, dass fast jedes Antibiotikum die pseudomembranöse Kolitis hervorrufen kann[22]. So entwickelte sich *C. difficile* schnell zum häufigsten Erreger von im Krankenhaus erworbenen Durchfallerkrankungen und ist für 95 % der Fälle von pseudomembranöse Kolitis verantwortlich[23].

C. difficile wird vergleichsweise selten im Darm von Erwachsenen nachgewiesen ($\leq$5 %), wobei dieser Anteil bei Aufnahme in ein Krankenhaus auf 20-40 % ansteigt. Dabei zeigt der Großteil der Patienten keine Symptome einer *Clostridoides difficile* Infektion (CDI) und gilt somit als asymptomatisch. Dies trifft auch auf die meisten Kleinkinder zu, von denen bis zu 80 % durch *C. difficile* im Darm besiedelt sind[14]. Die Kolonisation von Kleinkindern erfolgt meist bei der Geburt oder im ersten Lebensjahr, wobei der Anteil der kolonisierten Kinder mit steigendem Alter drastisch abnimmt. Allerdings häufen sich in letzter Zeit Berichte über das Vorkommen von pathogenen Stämmen in Säuglingen, wodurch eine zusätzliche Ansteckungsgefahr für Erwachsene entsteht[24]. *C. difficile* kann sich an diverse Umgebungen anpassen und sowohl auf Oberflächen im Krankenhaus als auch in der Natur überleben, wodurch etliche Quellen die Gefahr einer Transmission und einer möglichen Infektion durch das Bakterium bieten. Auch asymptomatische Patienten können, wenn auch im geringeren Maße als symptomatische Patienten, die Umwelt mit Sporen kontaminieren und so als potentielle Ansteckungsgefahr angesehen werden[25].
In den USA wurden laut einer Studie aus dem Jahre 2014 12,1 % der nosokomialen Infektionen durch *C.*

difficile ausgelöst[26]. In Deutschland beläuft sich der Anteil auf 10 % und stellt somit die vierthäufigste nosokomiale Infektionsart dar[14]. Das gehäufte Aufkommen und die Schwere der Krankheit belasten das Gesundheitssystem durch die erhöhten Kosten aufgrund längerer Patientenaufenthalte und schärferen Hygienemaßnahmen teilweise schwer[27]. So verursachen zum Beispiel wiederkehrende CDI, die in 20-30 % der Fälle vorkommen, geschätzte 2.8 Milliarden US Dollar zusätzliche Kosten pro Jahr in den USA[28].

1.1.1 Phänotypische und genotypische Eigenschaften

Um als anaerobes Bakterium in einer sauerstoffhaltigen Atmosphäre überdauern zu können, verkapseln sich *C. difficile* Bakterien zu Endosporen, die resistent gegen Hitze, Sauerstoff und Trockenheit sind. Dadurch sind sie besonders widerstandsfähig gegen äußere Einflüsse und lassen sich neben dem Darmtrakt von Mensch und Tier auch vermehrt in der Umwelt wie zum Beispiel in Böden, Abflüssen und Oberflächenwasser nachweisen[29]. Die Bildung der Endosporen machen das Bakterium zudem resistent gegen Desinfektionsmittel und erhöhen so das Risiko der Kontamination von beispielsweise Oberflächen in Krankenhäusern[22].
Diese phänotypischen Eigenschaften spiegeln sich auch in dem Genom von *C. difficile* wider, das bis zu 42 % größer ist als die Genome von anderen *Clostridien*-Spezies. Die Größe des Genoms und die Tatsache, dass viele kodierende Bereiche mit der Adaption und dem starken Ausbreiten im Verdauungstrakt assoziiert werden, unterstützen die Annahme, das *C. difficile* in unterschiedlichen Wirten über eine lange Zeit existieren und suboptimale Umgebungen durch die Bildung von Endosporen überstehen kann[22], [30]. Das Genom gilt weiterhin als hoch flexibel und besteht zu einem großen Teil aus mobilen genetischen Elementen wie Plasmiden und Phagen. Nur 16 % des Genoms werden zu dem Kerngenom, also zu dem in der Spezies konserviert vorkommenden Bereich, gezählt, was unter allen bakteriellen Spezies als einer der niedrigsten Werte gilt[31]. Während zu den Prophagen schon detailliertere Studien vorliegen, ist das Vorkommen von Plasmiden zwar bekannt, aber wenig untersucht. Prophagen beteiligen sich nachweisbar an der Regulation der Toxinproduktion, allerdings werden keine Virulenzgene über diese mobilen genetischen Elemente in das Bakterium mit eingebracht[27]. Die Virulenzfaktoren Enterotoxin A und Cytotoxin B befinden sich auf dem Pathogen-Locus „PaLoc“, der in allen toxischen *C. difficile* Stämmen vorkommt und eine entscheidende Rolle in der Pathogenese des Bakteriums spielt.

1.1.2 Pathogenese

Die Aufnahme des Erregers erfolgt fäkal-oral, wobei dieser meist in Form von Endosporen aus der Umwelt aufgenommen wird[32]. Nach der Aufnahme der Sporen ist eine gestörte Darmflora essenziell für die Ausbreitung des Organismus und der infektionsauslösenden Produktion der Toxine, wobei hier ein geschwächtes Immunsystem unterstützend wirken kann[27], [33]. Die mikrobielle Gemeinschaft im Darm wird besonders dann beschädigt, wenn Antibiotika verabreicht werden, die die im Darm natürlich vorkommenden Bakterien angreifen und gegen die *C. difficile* eine Resistenz zeigt. Hierzu werden vor allem Clindamycin, Penicillin, Ampillicin, Amoxicillin, Cephalsporine und für manche Stämme auch Chinolin-Antibiotika gezählt[22]. Durch das Verdrängen der anderen Bakterienarten können die *C. difficile* Sporen im Darm auskeimen und diesen übermäßig besiedeln. Als weitere Risikofaktoren für eine CDI werden neben einem fortgeschrittenen Alter der Patienten auch Chemotherapien, die den Darm angreifen, gastrointestinale Grunderkrankungen und chronische Nierenerkrankungen genannt[14], [34].
Als primäre Virulenzfaktoren gelten die von *C. difficile* hergestellten Toxine Enterotoxin A (tcdA) und Cytotoxin B (tcdB), wobei 20 % der Stämme ein zusätzliches binäres Toxin besitzen (CDT), das zunehmend mit schweren Krankheitsbildern in Verbindung gebracht wird, dessen genaue Wirkung aber noch genauer untersucht werden muss[35], [36]. Die Exprimierung beider oder auch nur eines der Toxine beschädigt die Darmzellen und stört die Salzaufnahme im Darm. Hierdurch kommt es zu einem großen Verlust an Flüssigkeit und Salzen, wodurch symptomatisch Durchfall bei dem Patienten auftritt. Zusätzlich können durch die entzündete Darmschleimhaut Fibrine ausgeschieden werden, die sich anschließend auf der

Wand des Kolons ablagern und eine so genannte Pseudomembran bilden, was die Erkrankung an einer pseudomembranösen Kolitis bedingt. Die Folgen können lebensgefährlich sein und zu einer Darmerweiterung (toxisches Megakolon), Darmverschluss (Ileus) oder Perforation des Kolons führen.

1.1.3 Diagnostik und Behandlung

Um endgültig als *C. difficile* Infektion diagnostiziert zu werden, muss neben klinischen Symptomen auch ein Labornachweis vorliegen. Zur mikrobiologischen Diagnostik kommt es allerdings nur, wenn der Patient ein fortgeschrittenes Alter hat, sich einer Antibiotikatherapie unterzieht und eines oder mehrere der folgenden Symptome zeigt:

- Abruptes Einsetzen von wässriger Diarrhö mit charakteristischem, fauligem Geruch
- Schmerzen im unteren Bauchbereich
- Fieber
- Länger als drei Tage anhaltende Diarrhö ohne das eine Verbindung zu einem anderen Krankheitserreger hergestellt werden kann (auch ohne Antibiotikatherapie)

Im Zuge der mikrobiologischen Diagnostik erfolgt der Nachweis der Toxingene tcdA und tcdB durch kommerzielle Enzymimmunoassays (EIA), wobei hier der Nachweis eines der beiden ausreichend ist. Für pathogene *C. difficile* Stämme, die nur das binäre Toxin enthalten gibt es noch keine routinierte Diagnostik, da diese noch als klinisch irrelevant gelten[37].
In 15-23 % der CDI Fälle können die Symptome, primär die Diarrhö, durch Beenden der Antibiotikatherapie innerhalb von 2-3 Tagen gestoppt werden. Kann die laufende Antibiotikabehandlung aus klinischer Sicht nicht unterbrochen werden, so wird eine CDI durch die Gabe der Antibiotika Metronidazol oder Vancomycin behandelt. Dies ist auch der Fall bei einem besonders schweren Verlauf der Krankheit oder bei fortgeschrittenem Alter des Patienten. Bei der Verabreichung von Metronidazol und Vancomycin ist zu beachten, dass diese eine schädigende Wirkung auf die Darmflora haben und somit den Patienten anfällig für wiederkehrende CDI machen[38]. Eine rezidivierende CDI wurde bei 20-25 % der Patienten, die mit Metronidazol oder Vancomycin behandelt wurden, nachgewiesen, wobei dies durch die Verabreichung des neu aufkommenden Antibiotikums Fidaxomycin nach ersten Erkenntnissen vermieden werden könnte[39], [40]. Eine wiederkehrende CDI, die durch den gleichen Stamm wie die initiale Infektion verursacht wird, von einer Neuinfektion durch einen neu aufgenommenen Stamm zu differenzieren stellt die Diagnostik vor eine große Herausforderung. Dabei hat die Unterscheidung einen großen Einfluss auf Studien, die die Wirksamkeit von Behandlungen und die Surveillance beurteilen[41].
Ein weiterer Therapieansatz ist die Stuhltransplantation (*faecal microbiota transplantation* (FMT)), wodurch die natürliche Darmflora wiederhergestellt wird und wiederkehrende Infektionen so gut wie ausgeschlossen werden können[42]. Alternative Methoden wie zum Beispiel eine Phagentherapie könnten rezidivierende CDI vermeiden, sind aber noch nicht bis zur Anwendung hin erforscht[27]. Bei besonders schwerem Verlauf und nicht erfolgreicher Therapie ist ein chirurgisches Eingreifen von Nöten.

1.1.4 Surveillance

Unter dem Begriff der „Surveillance“ von nosokomialen Infektionen versteht man „die fortlaufende, systematische Erfassung, Analyse und Interpretation der Daten zu diesen Infektionen, die zur Planung, Einführung und Evaluation von medizinischen Maßnahmen notwendig sind“[43]. Seit 2001 ist die Surveillance in Deutschland durch das Infektionsschutzgesetz (§23 Abs. 4 (IfSG)) geregelt[44]. Demnach müssen Leiter von Krankenhäusern oder ähnlichen Gesundheitszentren vermutete beziehungsweise bestätigte Fälle und/oder Tod durch bestimmte Infektionskrankheiten, sowie die Detektion eines bestimmten Pathogens an lokale Gesundheitsämter melden. Für eine umfangreichere und fachgerechtere Erfassung von meldungspflichtigen

Fällen wurde ein Netzwerk aus Nationalen Referenzzentren (NRZ) und Konsiliarlaboren (KL) geschaffen, die im engen Austausch mit dem Robert Koch-Institut (RKI), der nationalen Bundesoberbehörde für Infektionskrankheiten und nicht übertragbare Krankheiten, stehen[44]. Dabei übernehmen die Referenzzentren die Diagnostik und Feintypisierung sowie molekularbiologische Untersuchungen des Erregers, die epidemiologische Zusammenhänge aufdecken sollen. Fachlich unterstützt werden sie dabei von den Konsiliarlaboren, die primär als Beratungsstelle von betroffenen Patienten oder Krankenhausmitarbeitern fungieren.
Nach erlangter Kenntnis einer nosokomialen Infektion müssen Mitarbeiter der betroffenen Krankenhausstation den Fall innerhalb von 24 Stunden an das zuständige Gesundheitsamt melden. Wenn der gemeldete Fall der in §11 Abs. 2 des Infektionsschutzgesetzes festgehaltenen Falldefinition entspricht, leitet das Gesundheitsamt die Meldung an das NRZ weiter. Um die gemeldeten Fälle besser zwischen verschiedenen Stationen, Abteilungen und nicht zuletzt Krankenhäusern vergleichen zu können, entwickelte das NRZ 1996 das „Krankenhaus-Infektions-Surveillance-System“(KISS). Krankenhäuser können sich hier zur Erfassung und zum Abgleich von nosokomialen Infektionen je nach Bedarf für ausgewählte Module des KISS anmelden, die sich auf einen speziellen Risikobereich fokussieren (zum Beispiel Patienten auf Intensivstation oder Modul für Methilicin-resistente *Staphylococcus aureus*). Die dadurch erhobenen Daten werden dem NRZ regelmäßig in anonymisierter Form übermittelt und anschließend dort analysiert. Weiterhin werden bestätigte Fälle an das RKI weitergeleitet, wo die gesammelten Fallzahlen zentral überwacht und epidemiologisch ausgewertet werden[44]. Dadurch soll das gehäufte Vorkommen eines Pathogens zeitnah erkannt und zur frühzeitigen Detektion von Ausbrüchen führen, um rechtzeitig entsprechende Handlungen vornehmen zu können und so eine mögliche Epidemie zu vermeiden[45].
Die NRZ in Deutschland stehen weiterhin im engen Austausch mit internationalen Referenzlaboren und tragen auf den jeweiligen Falldefinitionen basierende Daten weiter auf die internationale Ebene, um so auch länderübergreifende Zusammenhänge zu erfassen. Zusammengetragen werden diese Daten durch das Europäische Zentrum für die Prävention und die Kontrolle von Krankheiten (ECDC), dass die Surveillance in der EU durchführt[46].

Das Robert Koch-Institut empfiehlt jeder Gesundheitseinrichtung ein System zur Überwachung der Inzidenz und des klinischen Verlaufs von *C. difficile* Infektionen. Laut Infektionsschutzgesetz (§6 Abs. 1 Nr. 5a (IfSG)) muss jeder CDI-Fall, der der Falldefinition entspricht und positiv auf mindestens eines der Toxingene tcdA und tcdB getestet wurde, gemeldet werden. Eine Verpflichtung der Stammtypisierung besteht allerdings nicht. Weitere Analysen zur Typisierung werden meist erst vorgenommen, wenn epidemiologische Verbindungen zwischen Fällen hergestellt werden können[47]. In Deutschland wird in den Referenzzentren primär die PCR Ribotypisierung zur Typisierung von *C. difficile* Isolaten eingesetzt, wobei bei einem Verdacht auf einen Ausbruch auch Ganzgenomsequenzierungen zur Detektion von epidemiologischen Zusammenhängen durchgeführt werden[14].

1.2 Typisierungsmethoden für *C. difficile* Isolate

Um Ausbreitungsrouten aufdecken und Quellen einer Infektion identifizieren zu können, werden oft molekulare Typisierungsmethoden herangezogen, die Isolate von pathogenen Bakterien spezifischen Phäno- oder Genotypen zuordnen[45]. Dabei kann der Anwendungsbereich der Typisierungsmethode unterschiedlich sein und von lokalen Untersuchungen im Krankenhaus über regionale Erfassungen in Referenzlaboratorien bis hin zur Analyse globaler Ausbreitungsszenarien, die durch internationale Kooperationen realisiert werden, reichen. Bis zu den 1990er Jahren waren hierfür nur phänotypische Methoden wie die Serotypisierung mittels Objektträger-Agglutination oder radio PAGE (Autoradiographie Polyacrylamid Gelelektrophorese) verfügbar[48]. Diese Methoden eigneten sich aufgrund ihres begrenzten Auflösungsvermögens und der geringen Reproduzierbarkeit sowie Typisierbarkeit jedoch nicht für epidemiologische Studien[49]. Mit der Entwicklung von genotypischen Methoden konnten diese

Kriterien deutlich verbessert und detaillierte Auflösungen erreicht werden. Genotypische Methoden lassen sich in bandenbasierte und sequenzbasierte Methoden unterteilen. Bei den bandenbasierten Methoden sind als meist verwendete Methoden die REA (Restriktions-Endonuklease Analyse), PFGE (*Pulsed-Field*-Gelelektrophorese), PCR (Polymerase-Kettenreaktion) Ribotypisierung und MLVA (Multi-Locus-Variable-Nummer-Tandemwiederholungs-Analyse) zu nennen, unter den sequenzbasierten Methoden wird primär die Multi-Locus-Sequenz-Typisierung (MLST) und die SNP (*Single Nucleotide Polymorphism*)-Analyse verwendet[48], [50], [51].

Da diese Methoden lange Zeit als zu teuer und zeitaufwendig galten, wurde 2010 die in der Proteomik schon länger verwendete MALDI-Methode (Matrixunterstützte Laserdesorption/Ionisation) mit der Massenspektroskopie (MS) kombiniert und zur MALDI-*Time-Of-Flight*(TOF) MS Methode weiterentwickelt. Anhand der generierten proteomischen Massenspektroskopie-Profile ermöglichte die Methode die Identifizierung von Mikroorganismen[52]. Allerdings wurden die Signale in dem resultierenden Massenspektroskopie-Profil oft durch nicht bekannte Proteine erzeugt, wodurch die Signale nicht zugeordnet und die Organismen nicht identifiziert werden konnten. Dieses Problem konnte mit der Proteotypisierung zumindest teilweise behoben werden[53]. Hier werden die massenspektroskopischen Signale in dem MALDI-TOF Spektrum, die von bekannten ribosomalen Proteinen verursacht werden, erfasst und aufgrund der Verschiebungen in dem Spektrum einer Aminosäuresequenz zugeordnet. Hierfür wird ein Aminosäurenkatalog von Isoformen von Allelen, demnach von nicht synonymen Mutationen in Genen, die für ribosomale Proteine kodieren, verwendet. Dadurch kann einem Isolat ein spezifischer von der Proteotypsierung abgeleiteter Typ zugeordnet werden.

Für die molekulare Typisierung von *C. difficile* Isolaten wird primär die PCR Ribotypisierung verwendet. Die Zuordnung eines endemischen *C. difficile* Stammes zu einem PCR Ribotyp wird universell eingesetzt und hat sich in der molekularen Epidemiologie etabliert[54], [55].

1.2.1 Die PCR-Ribotypisierung

Durch die PCR Ribotypisierung von *C. difficile* Isolaten werden regionale und länderübergreifende Verbreitungsmuster des Erregers erfasst und somit die Aufklärung von Ausbruchsgeschehen und möglichen Übertragungswegen ermöglicht[16], [56]. Die Methode basiert auf der Charakterisierung der zwischen den hoch konservierten 16S rRNA und 23S rRNA liegenden ITS-Regionen (*internal transcribed spacer*), die zusammen ein rRNA-Operon ergeben. Wie die meisten Prokaryoten besitzt auch das Genom von *C. difficile* multiple rRNA-Operons (zwischen sieben und 15), deren ITS-Regionen sich in ihren Längen unterscheiden[57], [58]. Dadurch werden bei einer PCR, bei der Primer für die konservierten 16S und 23S rRNA Regionen verwendet werden, DNA-Fragmente verschiedenster Längen amplifiziert. Die anschließende Gelelektrophorese trennt die DNA-Fragmente nach Größe auf und erzeugt so individuelle Bandenmuster, denen spezifische PCR Ribotypen zugewiesen werden können. Jedoch können die resultierenden Bandenmuster von unterschiedlichen Laboratorien voneinander abweichen, was die Reproduzierbarkeit der Methode und Interpretation der Ergebnisse erschwert. Diese Aspekte konnten mit der Entwicklung der Kapillar-Gelelektrophorese (CE) PCR Ribotypisierung stark verbessert werden[54]. Das neue Verfahren reduzierte die Anzahl der PCR Zyklen und ermöglichte die gleichzeitige Analyse einer höheren Anzahl von Isolaten. Zudem zeigten PCR Ribotypen, deren Bandenmuster vorher schwer voneinander zu unterscheiden waren, deutlich zu differenzierende Bandenmuster. Als Beispiel seien hier die PCR Ribotypen 014 und 020 genannt[54]. Mit der Entwicklung eines standardisierten und reproduzierbaren Protokolls wurde die CE PCR Ribotypisierung weiter verbessert und universell anwendbar[59]. Dennoch blieb das Problem der multiplen Nomenklaturen bestehen, denn die einst von Stubbs et al. etablierte Datenbank von PCR Ribotypen umfasst zwar über 650 Typen, ist aber nicht frei verfügbar[59], [60]. Aufgrund dessen entwickelten sich lokal eigenständige Nomenklaturen, die den Vergleich zwischen Laboratorien erschwerten.

1.2.2 Sequenzbasierte Typisierungsmethoden

Die Entwicklung der Multi-Locus-Sequenz-Typisierung (MLST) eröffnete eine gut aufgelöste Möglichkeit Bakterien anhand ihrer sieben Haushaltsgene zu genotypisieren. Bakterien konnten so innerhalb ihrer Spezies anhand von Nukleotidsequenzen in spezifischere Kladen unterteilt werden, was einen bis dato detaillierteren Einblick in mögliche Transmissionswege, insbesondere von Pathogenen, ermöglichte. Dafür werden in der MLST Allelprofile für jedes Isolat erstellt, wobei jeder individuellen Nukleotidsequenz, die sich an den Koordinaten der Loci im MLST Schema befindet, eine einzigartige Allelnummer zugeordnet wird. Dabei steht ein Allelunterschied für ein genetisches Ereignis, ungeachtet der tatsächlichen Anzahl an Punktmutationen[61] (Abbildung 1.1). Das dadurch entstehende Allelprofil wird anschließend einem MLST-Sequenztypen zugewiesen. Um einen standardisierten und reproduzierbaren Vergleich von Bakterien aus aller Welt zu ermöglichen wurde die frei verfügbare Softwareplattform BigsDB entwickelt. Diese stellt neben weiteren Informationen wie antibiotikaresistente Gene oder Stoffwechselwege auch kuratierte Datenbanken der MLST-Sequenztypen für einzelne Bakterienspezies zur Verfügung[62].
Mit dem Aufkommen der Sequenziermethoden der nächsten Generation (*Next Generation Sequencing*) stieg die Anzahl der verfügbaren bakteriellen Genomsequenzen, die zunehmend die Breite der ganzen Domäne repräsentierten. Infolgedessen wurde ein Genotypisierungsschema entwickelt, dass vergleichbare Analysen zwischen allen Bakterienspezies ermöglicht und eine Genomsequenz bis zum bakteriellen Stamm zuordnen konnte. Hierfür wurden 53 ribosomale proteinkodierende Gene bestimmt, welche konserviert in der ganzen Domäne vorkommen[63]. Das so genannte ribosomale MLST Schema (rMLST) wurde fortan zur Typisierung und taxonomischen Zuordnung bakterieller Sequenzdaten verwendet. Allerdings deutet ein vermehrtes Aufkommen eines rMLST-Typen, genauso wie eines PCR Ribotypen nur auf einen Ausbruch hin, eine genaue Auflösung der Transmissionswege war mit diesen Methoden nicht möglich.
Je weiter sich die Sequenziertechniken entwickelten und bald schon Hochdurchsatz-Sequenzierung zuließen, desto vielfältiger wurde auch die Entwicklung der bioinformatischen Werkzeuge, um die entstehenden Sequenzdaten zu verarbeiten. Diese bioinformatischen Werkzeuge ließen bald die Detektion von Punktmutationen zwischen bakteriellen Genomsequenzen zu. Die so genannte SNP-Analyse wurde das Standard-Werkzeug zur Detektion von nahen genomischen Verwandtschaften zwischen Bakterien und somit zur Detektion von Transmissionswegen[64]. Da diese Analyse allerdings rechenintensiv und somit zeitlich aufwendig war und umfassendes bioinformatisches Wissen erfordert, war das Interesse an alternativen Analysemethoden groß. Es galt die zuvor universell anwendbare rMLST wieder auf Speziesebene zu bringen und dadurch eine bessere Auflösung zu gewährleisten.

Kerngenom- und Ganzgenom-MLST

War für die vorangegangenen MLST Schemata BIGSdb die zentrale Plattform, so entwickelten jetzt unabhängig voneinander verschiedene Arbeitsgruppen Kerngenom-MLST (cgMLST) Schemata für ausgewählte Bakterien. Dabei hat die cgMLST zum Ziel, ein Set an nicht-repetitiven Genen zu umfassen, die konserviert in allen Stämmen der Spezies vorkommen. Für *C. difficile* stehen momentan drei cgMLST Schemata zur Verfügung, zwei davon werden von kommerziellen Softwares angeboten und eines durch die online frei verfügbare Softwareplattform EnteroBase. Das in EnteroBase zugängliche cgMLST Schema ist eine Teilmenge eines Ganzgenom-MLST (wgMLST) Schemas, das alle Einzelkopie Orthologe des Pangenoms, das aus einem repräsentativen Set von 442 *C. difficile* Genomen gebildet wurde, umfasst[65]. Für das cgMLST Schema wurden aus den 11.490 bestimmten Genen diese ausgewählt, die in $\geq$98 % der Referenzgenome vorhanden sind und in $\geq$94 % als intakte Gene vorkommen. Daraus ergab sich ein cgMLST Schema mit 2.556 Loci. BioNumerics stellt neben dem cgMLST Schema auch ein wgMLST Schema zur Verfügung. Hier wurde für 259 Referenzgenome anhand eines Multi-Reziprok-BLAST-Verfahrens 6.713 Loci bestimmt, die das akzessorische Genom der Referenzgenome repräsentieren [66]. Zusammen mit dem 1.999 Loci umfassenden cgMLST Schema ergibt sich hier ein wgMLST Schema mit 8.712 Loci.
In SeqSphere$^+$ wird ein cgMLST Schema mit 2.270 Loci angeboten, das durch einen genomweiten

gene-by-gene Vergleich von elf Referenzgenomen erstellt wurde [67].
Dadurch, dass die Unterschiedsberechnungen auf Allele begrenzt und nicht für jeden einzelnen Genunterschied erfolgt, können mit der cgMLST-Analyse eine Vielzahl an Isolate innerhalb kürzester Zeit in ihren Kerngenomen miteinander verglichen werden[68], [69] (Abbildung 1.1).

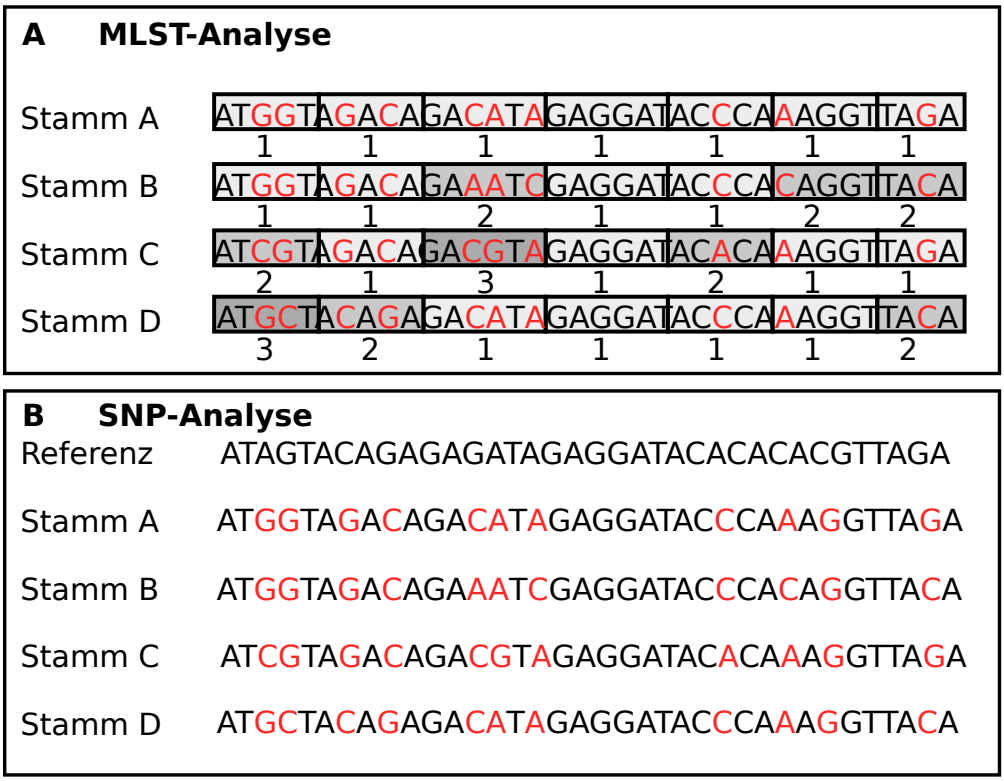

Abbildung 1.1: Erfassung von genomischen Unterschieden durch die MLST- und die SNP-Analyse an einem Beispiel von vier Sequenzvergleichen. **(A)** Bei der MLST-Analyse werden die Sequenzen in Abschnitte eingeteilt, so genannte Loci. Jeder individuell erfassten Sequenz an einer Locus-Stelle wird eine individuelle Allelnummer zugeordnet. Dadurch entsteht für jede Sequenz ein individuelles Allelprofil, anhand derer die vier Sequenzen miteinander verglichen werden können. **(B)** Bei der SNP-Analyse werden nach der Alignierung der Illumina-*Reads* gegen eine Referenzsequenz die von der Referenz abweichenden Basen detektiert (hier rot markiert). Bei einem Vergleich der Sequenzen wird jede Punktmutation als Unterschied gewertet.

Die SNP-Analyse

Die SNP-Analyse verspricht durch die Detektion einzelner Punktmutationen gegenüber einer Referenzsequenz ein hohes Auflösungsvermögen (Abbildung 1.1). Dafür werden entweder durch die Sequenzierung entstandene kurze Genomfragmente (*Short-reads* oder auch Illumina-*Reads*) oder durch das Überlappen dieser Genomfragmente gebildete längere zusammenhängende Sequenzfragmente (assemblierte *Contigs*) gegen ein Referenzgenom gemappt. Das bedeutet, dass für jedes Genomfragment die Stelle auf dem Referenzgenom gesucht wird, von der es am wahrscheinlichsten abstammt. Hierfür können verschiedene Algorithmen eingesetzt werden, wobei die meisten auf der so genannten *seed-and-extend* Heuristik beruhen[70]. Dabei werden Ähnlichkeiten zwischen Referenz und Illumina-*Read* über kurze Übereinstimmungen, so genannte *seeds*, gesucht. Diese *seeds* dienen als Ausgangspunkt für die weitere Alignierung des *Reads* an die Referenzsequenz. Als Referenz sollte hier ein vollständig sequenziertes Genom der gleichen Spezies gewählt werden, das im besten Fall nah verwandt mit dem zu untersuchenden Datensatz ist. Anschließend werden an jeder Stelle des Genoms die Baseninformationen aller alignierter Illumina-*Reads* zusammengefasst. So kann für jedes Set an Illumina-*Reads* eine Konsensussequenz erstellt und genomische Unterschiede zwischen den zu vergleichenden Isolaten detektiert werden. Dieser Prozess wird auch *Variant-calling* genannt. Die in vielen Publikationen zur Detektion von Transmissionswegen verwendete

SNP-Analyse [71]–[73] fasst die erstellten Konsensussequenzen in ein Alignment zusammen und filtert aus diesem die SNPs heraus. Anschließend werden aus diesem SNP-Alignment die Punktmutationen entfernt, die mutmaßlich auf rekombinanten Regionen des Genoms detektiert wurden. Bei einer Rekombination werden Fragmente eines Genoms durch Teile eines anderen Genoms oder freier DNA (zum Beispiel Plasmide) ersetzt [74]. Da die dadurch entstandenen Differenzen zwischen den zu vergleichenden Genomen nicht evolutionärer Natur sind, sollten diese in der SNP-Analyse ausgeschlossen werden, um die klonalen Beziehungen zwischen den Isolaten erfassen zu können. Anhand des dadurch gebildeten SNP-Alignments können dann genomische Zusammenhänge zwischen den Isolaten geschlussfolgert werden.
Für die Schritte des *Read-Mappings* und *Variant-callings* stehen eine Vielzahl an bioinformatischen Werkzeugen zur Verfügung, wodurch jede SNP-Analyse auf den zu analysierenden Datensatz angepasst werden kann[75], [76]. Allerdings erschwert dies auch die Standardisierung und Vergleichbarkeit der generierten Ergebnisse. Neben der manuellen Anwendung der Werkzeuge können SNP-Analysen auch durch bereitgestellte Pipelines durchgeführt werden, die eine Auswahl an Werkzeugen für die einzelnen Schritte zulassen[77] oder abhängig von dem Sequenzdatensatz eine eigene Wahl treffen[78].

1.3 Molekulare Epidemiologie von *C. difficile*

1.3.1 Übersicht über die Epidemiolgie der letzten Jahre

Durch die Entwicklung der sequenzbasierten Typisierungsmethoden konnte die Bewegung des Pathogens detaillierter erfasst werden und zeigte eine sich ständig wandelnde Epidemiologie von *C. difficile*. So sorgte zum Beispiel das Aufkommen des hypervirulenten, fluorchinolonresistenten und mit höheren Sterberaten verbundenen PCR Ribotyps RT027 in Nordamerika international für großes Aufsehen, da dieser Stamm in andere Länder eingebracht wurde und international Ausbrüche verursachte[15], [16], [27]. Auch in Deutschland erlangte CDI durch das Vorkommen des RT027 Stammes neue Bedeutung und wurde in den folgenden Jahren intensiver beobachtet[79]. In Australien wiederum trat der Stamm kaum in Erscheinung, da dort Fluorchinolone selten als Antibiotikum verabreicht werden[80].
Des Weiteren wurde das ursprünglich meist nosokomiale Infektionen verursachende Bakterium immer häufiger in der Gesellschaft nachgewiesen und betraf auch Menschen, die nicht dem allgemeinem Patientenbild einer CDI entsprachen[27]. Das Vorkommen von *C. difficile* Stämmen außerhalb von Krankenhäusern oder anderen Gesundheitseinrichtungen wurde anfangs in Bezug mit Lebensmittelvergiftungen gebracht, da sich Nachweise des eigentlich mit Nutztieren assoziierten PCR Ribotypen RT078 im Menschen häuften[81], [82]. Ein endgültiger Beweis eines direkten Übertragungsweges zwischen Lebensmittel und Mensch steht jedoch noch aus, wobei der Stamm erfolgreich auf Fleisch nachgewiesen wurde[83]–[85]. Zudem zeigten andere Studien genomisch nahe Verwandtschaften zwischen Mensch und Schwein und begründeten damit die globale Verteilung des RT078 Stammes[86], [87]. Das Vorkommen von *C. difficile* in Tieren ist allerdings nicht nur auf den RT078 Stamm beschränkt. In Hühnern auf Geflügelfarmen konnten bis zu zwölf unterschiedliche Ribotypen nachgewiesen werden, ohne das sich hierbei ein Stamm als dominant vortat[88], [89].
Neben dem Nachweis in Mensch und Tier wurden auch schon Isolate aus Abwässern[90], Salz- und Süßwasserquellen[91], Böden und anderen Umweltquellen[29], tierischem Dung[92] und sogar aus Haustieren isoliert, was das Vorkommen von *C. difficile* in der häuslichen Umgebung erklärt[93], [94]. Obwohl ein Rückgang an *C. difficile* Sporen bei Zunahme der Temperatur in Gülle auf Schweinekompostbasis gezeigt wurde, konnten in fertigen Kompost-Produkten noch Sporen nachgewiesen werden[95]. Eine Ausbringung des kontaminierten tierischen Dungs auf landwirtschaftlich genutzte Flächen stellt somit eine weitere potentielle Quelle für die Aufnahme von *C. difficile* Sporen durch den Menschen dar[95], [96].

Mit der Entwicklung der 7-Gen MLST wurde erstmals ein diverses Set an *C. difficile* Stämmen einer phylogenetischen Untersuchung unterzogen und dadurch in fünf distinkte Kladen eingeteilt[97]. Diese

Kladen wurden durch ganzgenombasierte Analysen durch weitere Genome ausgebaut und bestätigt[98]. Eine Assoziation der Kladen mit geographischer Herkunft oder Wirtsspezies wurde nicht festgestellt, jedoch bildeten sich hierfür mit der Zeit durch die Zunahme der Genome in den Kladen Tendenzen ab. So bestehen die Kladen 2-5 hauptsächlich aus Isolaten, die sich einem bestimmten Ribotypen zuordnen lassen, während Klade 1 hoch divers ist und viele verschiedene Ribotypen umfasst, die nosokomiale Infektionen auslösen können (zum Beispiel RT014, RT020, RT001). Isolate des anfangs schon erwähnten PCR Ribotypen 027 bilden den Hauptanteil der zweiten Klade. Dabei umfasst die Klade beide Untergruppen des Stammes, die die voneinander unabhängige Einbringung der Fluorchinolonresistenz in zwei unterschiedliche Linien demonstrieren und beide global auftraten (FQR1 und FQR2 [15]). Zusätzlich befinden sich in der zweiten Klade noch weitere toxische Ribotypen, die zumindest begrenzt auf einem Kontinent auftaten. So wurde der Ribotyp 244 hauptsächlich in Australien und Neuseeland gefunden[99], während sich RT176 in Europa ausbreitete[100]. PCR Ribotyp 023, der den Großteil der dritten Klade ausmacht, wurde bis jetzt nicht detaillierter in der Forschung behandelt. Dies traf anfangs auch auf PCR Ribotyp 017 zu, der die vierte Klade bildet. Die kontinuierliche Detektion dieses Stammes in Asien und das vermehrte Aufkommen in anderen Ländern machte den Ribotypen immer mehr zum Gegenstand weiterer Untersuchungen und galt sogar als Auslöser einer globalen Pandemie[101]–[103]. Isolate der Klade 5 zeigen eine starke geographische Assoziation mit Australien. Jedoch umfasst die Klade auch Isolate des zuvor schon erwähnten PCR Ribotypen 078, der sich über Landesgrenzen hinaus verbreitet[27].

Die meisten bekannten Isolate stammen aus dem Vereinigten Königreich, Australien und Nordamerika[27]. Während PCR Ribotypen wie 014, 020 und 002 universell und kontinuierlich in Krankenhäusern zu finden sind ergibt der Vergleich von umfangreichen Studien für manche *C. difficile*-Typen lokale Tendenzen. So wurden neben den erwähnten Ribotypen in Nordamerika vornehmlich Isolate der Ribotypen 001 sowie 053 nachgewiesen[104] und in Australien eher Ribotypen wie 054,056 und 070[105]. Für viele Länder liegen solch umfangreiche Studien allerdings nicht vor. Für Deutschland lässt sich anhand der Analyse von 16 Ausbrüchen ein vermehrtes Aufkommen der PCR Ribotypen 027 und 001 feststellen[106]. Publikationen aus Südamerika und Afrika beziehen sich wiederum meist nur auf Krankenhausstudien und lassen eine allgemeine Aussage über die Verbreitung bestimmter *C. difficile* Stämme nicht zu[27].

1.3.2 Detektion von Transmissionswegen durch genomische Analysen

Die SNP-Analyse hat sich in der molekularen Epidemiologie als Standardmethode zur Untersuchung von genomischen Beziehungen zwischen Isolaten, insbesondere zur Aufdeckung von möglichen Transmissionswegen, etabliert[107]. Dabei werden zwei Isolate zu einer Transmissionskette gezählt, wenn sie sich in weniger SNPs als ein festgesetzter Grenzwert unterscheiden[108]. Dieser Grenzwert wird für jedes Pathogen spezifisch berechnet, wobei hierfür meist ein auf der Mutationsrate des Organismus pro Jahr basierendes Evolutionsmodell verwendet wird[109], [110]. Es gibt jedoch kein einheitliches Verfahren um einen solchen Grenzwert zu bestimmen[108], [111].

Für *C. difficile* wird ein von Eyre et al. etablierter Grenzwert von $\leq$2 SNPs zur Detektion von Genomen verwendet, deren Isolate mit einer 95 %igen Wahrscheinlichkeit einer Transmissionskette angehören[72]. Ein gemeinsamer Ursprung der Infektion wird bei einer genomischen Distanz von >10 SNPs ausgeschlossen. Dieser Grenzwert wurde durch ein auf der Koaleszenztheorie basierendem komplexen Evolutionsmodell berechnet, das auf 90 Genome von Isolaten aus rezidivierenden Patienten angewendet wurde[72]. Dabei wurden Isolate des ersten und des letzten Aufenthalts der Patienten berücksichtigt und die SNP Distanz zwischen den zugehörigen Genomen bestimmt. Es wurde eine lineare Regression zwischen den SNP Distanzen und der Zeit, die zwischen den Isolationszeitpunkten lag, berechnet und anschließend für 200 Permutationen jedes Datenpunktes ein Prognoseintervall bestimmt. Aufgrund dessen konnte dann die Prognose für eine genomische Distanz von $\leq$2 SNPs zwischen Isolaten abgeleitet werden, die aus dem gleichen Patienten stammen. Das der so bestimmte Grenzwert auch auf Transmissionen zwischen Patienten, die gemeinsame Zeit auf einem Zimmer oder einer Station teilten, anwendbar ist, wurde durch eine groß angelegte Studie

in vier Krankenhäusern in Oxford zwischen 2006 und 2010 bestätigt. Die gleichen Grenzwerte können auch angewendet werden um ein Rezidiv von einer Neuinfektion zu unterscheiden[40].
Seit der Veröffentlichung der Publikation im Jahre 2013 hat sich der Grenzwert von $\leq$2 SNPs als gängiges Mittel zur Detektion von möglichen Transmissionsketten von *C. difficile* etabliert. So wurden sowohl lokale und regionale Ausbrüche von *C. difficile* aufgedeckt[55], [112] als auch Transmissionsketten zwischen verschiedenen Wirtsspezies (zum Beispiel Schwein und Mensch[87]) und länderübergreifende Zusammenhänge hergestellt[56].
Die Ganzgenomsequenzierung wird laut ECDC in der Hälfte der Länder in Europa als Mittel für die nationale Surveillance eingesetzt(Stand: 2016[113]). Um auch auf internationaler Ebene Stämme miteinander vergleichen zu können, ist eine universelle Nomenklatur und eine standardisierte Analyse, die auch von Forschern mit unterschiedlicher Expertise durchgeführt werden kann, essenziell.

1.4 Die Softwareplattform EnteroBase

Mit der rasanten Entwicklung der Sequenziertechniken und der dadurch stetig zunehmenden Anzahl an Sequenzdaten, die in öffentlichen Repositorien wie NCBI zur Verfügung stehen, galt es eine Plattform zu entwickeln, die für ausgewählte Mikroorganismen eine schnelle und unkomplizierte Analyse dieser Daten ermöglicht. Seit 2014 bietet die Website EnteroBase (`https://enterobase.warwick.ac.uk`) eine integrierte Softwareplattform, die automatisch *Short-read* Archive nach Illumina-*Reads* von *Salmonella*, *Escherichia*, *Yersinia*, *Moraxella*, *Vibrio*, *Heliobacter* und *Clostridioides* durchsucht. Zusätzlich ist das Hochladen eigener, nicht veröffentlichter Illumina-*Reads* durch Nutzer der Plattform möglich. Anschließend werden aus den *Short-reads* durch eine standardisierte Assemblierungs-Pipeline Entwürfe von Genomen gebildet, die aus einer Vielzahl an *Contigs* bestehen[65], [114]. Die dadurch entstandenen Assemblierungen werden dann zusammen mit den entsprechenden Metadaten in die Datenbank eingetragen. Hierbei wird bei manuellem Hochladen die Angabe des Kontaktlabors vorausgesetzt, um bei möglichem Interesse an einem Austausch der Daten die Kontaktaufnahme zu vereinfachen. Sensiblere Metadaten, die nicht in der Datenbank für jeden öffentlich sichtbar sein sollen, können in einem so genannten „*Custom View*“, der nur für den jeweiligen Nutzer zugänglich ist, den Einträgen in EnteroBase zugeordnet werden. Des Weiteren werden Genome, die zu demselben Stamm gehören, in einem so genannten „*Uberstrain*“ zusammengefasst[68]. Assemblierungen, die eine ausreichende Qualität zeigen, werden anschließend anhand der 7-Gen MLST und rMLST, sowie der Kerngenom- und Ganzgenom MLST Schemata (siehe Kapitel 1.2.2) genotypisiert (für *Moraxella*, *Vibrio*, *Heliobacter* ist nur das rMLST Schema verfügbar)[68], [114]. Die durch die cgMLST bestimmten Sequenztypen werden anschließend durch hierarchische Clusterung auf Basis der paarweisen Übereinstimmungen an Kerngenom-Allelen auf verschiedenen Ebenen in hierarchische Cluster (HC) zusammengefasst. Zudem können die *Contigs* der Assemblierungen gegen ein in der Datenbank verfügbares Komplettgenom gemappt und SNP Distanzen berechnet werden.
Die genomischen Zusammenhänge können in EnteroBase mit Hilfe des Visualisierungswerkzeugs GrapeTree dargestellt werden, wobei zwischen verschiedenen Algorithmen gewählt werden kann[115]. Durch die Verbindung von GrapeTree mit der Datenbank können die Endpunkte der berechneten Bäume beliebig nach verschiedenen Metadaten eingefärbt werden. Hierfür können in der Datenbank verfügbare oder Informationen aus dem selbst erstellten *Custom View* als Grundlage dienen.

1.5 Zielsetzung

Mit der Entwicklung der Softwareplattform EnteroBase und der dort implementierten Datenbank für *C. difficile* Genome ermöglicht sich erstmals die Analyse einer globalen Sammlung von insgesamt 13.515 Genomen. Der Mitaufbau und die Kuratierung dieser zunächst exklusiv verfügbaren Datenbank wurde für das Ziel der Erfassung der Populationsstruktur von *C. difficile* vorausgesetzt. Konnten bisherige Populationsstudien nur auf begrenzte Datensätze durchgeführt werden, so hatte diese Arbeit zum Ziel die Gesamtheit der verfügbaren Sequenzdaten von *C. difficile* zu betrachten und die in EnteroBase zugänglichen bioinformatischen Werkzeuge für eine standardisierte Analyse der Daten zu nutzen. Anhand dieser umfangreichen Datensammlung galt es eine global einheitlich durchführbare Typisierungsmethode zu ermitteln, die in Zukunft als Ersatz für die routinemäßig verwendete PCR Ribotypisierung angesehen werden könnte.
Um die so erlangten Aussagen mit den in der Literatur getätigten in Relation setzen zu können, war es vom großen Interesse erstmals einen quantitativen Vergleich zwischen der standardmäßig für genomische Untersuchungen verwendete SNP-Analyse und der cgMLST-Analyse in EnteroBase anzustellen. Die Übereinstimmung der durch die beiden Methoden berechneten genomischen Distanzen ist besonders bei der Detektion von Ausbrüchen von großer Bedeutung, da hier ein enger Grenzwert von $\leq$2 SNPs angelegt wird.
Um zu demonstrieren, dass die cgMLST-Analyse eine Alternative zur SNP-Analyse darstellt, wurde anhand dieser eine retrospektive Ausbruchsdetektion in einem Netzwerk von Krankenhäusern, sowie eine Reanalyse von vier publizierten *C. difficile* Ausbrüchen durchgeführt. Neben klinischen Isolaten galt es auch genomische Zusammenhänge zwischen Umweltproben zu untersuchen. Da erste Ergebnisse nahe genomische Verwandtschaften zwischen epidemiologisch unabhängigen Isolaten schlussfolgern ließen, wurde dieses Phänomen im weiteren Verlauf der Arbeit näher untersucht, wobei erneut der Umfang der Datenbank genutzt werden konnte. Dabei taten sich sowohl Vor- als auch Nachteile der SNP- und cgMLST-Analyse hervor, die für ein besseres Verständnis der verwendeten Methoden in dieser Arbeit ausführlich beschrieben sind und in Zukunft von Nutzern berücksichtigt werden sollten.

Das übergeordnete Ziel dieser Arbeit war es, die bioinformatischen Werkzeuge und die *C. difficile* Datenbank in EnteroBase zu verwenden um sowohl Typisierungen der Genome des Bakteriums als auch detailliertere Ausbruchsanalysen in einem globalen Kontext vorzunehmen und zu demonstrieren, dass diese Sachverhalte in Zukunft mit Hilfe von EnteroBase durch Wissenschaftler ohne bioinformatischem Hintergrund untersucht werden können.

Kapitel 2

Methodik

Die in dieser Arbeit eingesetzte Methodik umfasst hauptsächlich computergestützte Analysen. Um Sequenzdaten zu verarbeiten, wurden zum einen von Kollegen des Leibniz-Instituts DSMZ in Braunschweig implementierte, zum anderen selbst erstellte Kommandozeilen-Skripte (Shell) für unixartige Betriebssysteme verwendet. Dadurch wurde die Anwendung verschiedenster bioinformatischer Applikationen ermöglicht und die Reproduzierbarkeit der Analysen sichergestellt. Die anschließende Auswertung und Anfertigung von Abbildungen erfolgte in der Programmiersprache R [116]. Auch hier wurde die Reproduzierbarkeit und universelle Anwendung von sich wiederholenden Analyseschritten auf verschiedene Datensätze durch Erstellung von Skripten ermöglicht und mit R Markdown dokumentiert. Mit R erstellte Funktionen, mit der die Datensätze prozessiert wurden, sind auf der Plattform Github zur allgemeinen Verwendung verfügbar. Eine weitere Verarbeitung der Sequenzdaten erfolgte durch implementierte Werkzeuge in der Software-Umgebung EnteroBase. Die in der EnteroBase Umgebung zugängliche *Clostridioides difficile* Datenbank wurde im Zuge dieser Arbeit mit aufgebaut und kuratiert, sodass die Datenbank zur Analyse der Populationsstruktur von *C. difficile* genutzt werden konnte.
Zur Untersuchung epidemiologischer Fragestellungen wurden schwerpunktmäßig drei Datensätze analysiert. Um die Anzahl der Einträge in der *C. difficile* Datenbank mit bekannten PCR Ribotypen zu erhöhen wurden im Laufe dieser Arbeit ergänzend weitere Genome von Isolaten, für die diese Information vorlag, sequenziert. Da es in dieser Arbeit auch publizierte Ergebnisse zu rekonstruieren galt um diese mit in EnteroBase generierten Ergebnissen zu vergleichen, wurden zusätzlich eine Vielzahl an publizierten Datensätzen untersucht. Durch den Zugriff auf die umfangreiche Datenbank konnten die Analysen auf 13.515 genomische Sequenzen von *C. difficile* ausgeweitet werden.

2.1 Generierung der Sequenzdaten

2.1.1 Stammsammlung

Um die Ausbreitung von *C. difficile* Bakterien untersuchen zu können, wurden die im Folgenden beschriebenen Stammsammlungen sequenziert. Alle Isolate wurden wie in Kapitel 2.1.3 beschrieben prozessiert. Eine detailliertere Beschreibung der daraus resultierenden Datensätze ist in Kapitel 2.2.1 zu finden.

Isolate von Patienten mit wiederkehrender *C. difficile* Infektion

Um wiederkehrende *Clostridioides difficile* Infektionen (CDI) zu untersuchen wurden insgesamt 183 *C. difficile* Isolate in Kooperation mit Dr. Lutz von Müller (Konsiliarlabor *Clostridioides difficile*, Saarland) gesammelt. Hier wurden zunächst Primärausstriche von Stuhlproben von Patienten mit CDI auf Kulturagarplatten bei 4°C für fünf Monate aufbewahrt, um bei einer möglichen Reinfektion des Patienten eine Untersuchung von Isolaten aus beiden Episoden zu ermöglichen. Die Platten wurden an das Leibniz-Institut

DSMZ geschickt. Dort wurden durch Vera Junker (Leibniz-Institut DSMZ, Braunschweig) von beiden Platten so viele Isolate wie möglich gepickt und anschließend kultiviert.

Isolate aus einem Netzwerk von Krankenhäusern

In Kooperation mit Prof. Dr. Uwe Groß (Institut für medizinische Mikrobiologie, Göttingen), Leiter eines Diagnostiklabors, welches Services in der medizinischen Mikrobiologie für Krankenhäuser in Deutschland anbietet, wurden insgesamt 309 *C. difficile* Isolate gesammelt. Um eine möglichst repräsentative Sammlung zu erhalten, wurden in fünf Krankenhäusern, die regulär Patienten miteinander austauschen, über drei Jahre systematisch *C. difficile* Isolate von Patienten, die an einer CDI erkrankt waren, isoliert. Eines der Krankenhäuser stieg nach den ersten beiden Beprobungs-Perioden aus der Studie aus. Die Studie wurde als Ersatz an einem sechten Krankenhaus weitergeführt. Dafür wurden in jedem Jahr die ersten 20 CDI Fälle beprobt. Die Isolate wurden durch Dr. Alexander Indra und Marion Blaschitz (Österreichische Agentur für Gesundheit und Ernährungssicherheit (AGES), Wien) PCR ribotypisiert. Die epidemiologischen Informationen über den Zeitpunkt der *C. difficile* Infektion und das Krankenhaus beziehungsweise Station, auf der der Patient hospitalisiert war, wurden von Ortrud Zimmermann (Diagnostiklabor des Instituts für medizinische Mikrobiologie, Göttingen) anonymisiert zur Verfügung gestellt.

SOARiAL Isolate

In dem Projekt SOARiAL wurde die mögliche Übertragung von antibiotikaresistenten Keimen in Dünger durch dessen Ausbringung auf landwirtschaftlich genutzte Flächen in den Boden untersucht. Zusätzlich galt es die durch die Einbringung des Dungs entstehenden Staubwolken auf Bakterienbelastung zu testen. Während des Projekts wurden ein Feldversuch in Brandenburg und mehrere Windkanalversuche an dem Leibniz-Zentrum für Agrarlandschaftsforschung (ZALF) in Müncheberg durchgeführt. Dafür wurde zunächst die Bakterienbelastung von Sammelkotproben, die in einer Geflügelmastfarm in Brandenburg gesammelt wurden, untersucht, um so geeigneten Dung für die Versuche zu finden. Dieser Dung wurde auf einem Testfeld in Friedrichshof-Müncheberg durch einen landwirtschaftlichen Düngerstreuer ausgebracht und anschließend mit einem Feldgrubber in den Boden eingearbeitet. Vor und nach Ausbringung des Dungs wurden an drei repräsentativen Stellen Bodenproben genommen. Die gleichen Stellen wurden 2, 4, 7, 10, 14 und 19 Wochen nach der Ausbringung erneut beprobt. Für den Windkanalversuch wurde Dung aus der gleichen Geflügelmastfarm manuell in landwirtschaftlichen Boden eingearbeitet und anschließend in den Windkanal gefüllt. Es wurden drei Windgeschwindigkeiten angewendet und der dadurch aufgewirbelte Staub wurde durch ein Aerosolsammelsystem, welches auftreffende Partikel in phosphatgepufferte Salzlösung (PBS) leitet, eingefangen. Alle gesammelten Proben wurden bei 4°C aufbewahrt und innerhalb von 24 Stunden auf ihren mikrobiellen Gehalt analysiert. Dabei konnten aus allen Proben insgesamt 209 *C. difficile* Isolate isoliert werden. Der Feldversuch und die Windkanalversuche wurden von Dr. Nadine Thiel (ehemals Leibniz-Institut DSMZ, Braunschweig) in Zusammenarbeit mit den Kooperationspartnern von den Leibniz-Instituten ZALF, ATB, TROPOS und der FU Berlin und die anschließenden Analysen im Labor mit der Unterstützung von Vera Junker durchgeführt. Detailliertere Angaben zu dem Versuchsaufbau und der Durchführung sind in der Publikation von Thiel et al. [117] beschrieben.

2.1.2 PCR Ribotypisierung und Literaturrecherche

Insgesamt wurden im Rahmen der vorliegenden Arbeit 656 Genome von PCR ribotypisierten Isolaten sequenziert und zur *C. difficile* Datenbank in EnteroBase hinzugefügt. Um die Datenbank so gut wie möglich nutzen zu können, wurden für jeden der 13.515 Einträge öffentlich verfügbare Metadaten, darunter auch der PCR Ribotyp, ergänzt. Dafür wurden die Einträge der Datenbank nach ihren verlinkten Projekten aufgeteilt und geprüft, ob zu diesem Projekt eine Publikation vorliegt. Wenn dies der Fall war, wurde die Publikation auf zugehörige Metadaten für die jeweiligen Einträge durchsucht. Dabei wurde darauf geachtet, dass nur bei eindeutigem Verweis auf die dem Eintrag zugehörigen Illumina-*Reads* (*Accession number*) die Information

in die Datenbank überführt wurde. Dadurch konnte für 1.607 Einträge die Information des PCR Ribotypen ergänzt werden. Insgesamt beinhaltete die Datenbank 2.263 PCR ribotypisierte Einträge.

2.1.3 DNA Extraktion und Sequenzierung

Die DNA Extraktion und Vorbereitung der DNA zur Sequenzierung der zuvor beschriebenen Proben wurde zu großen Teilen von Vera Junker durchgeführt. Die DNA Extraktion wurde mit dem DNeasy Blood & Tissue Kit von Qiagen® wie im Handbuch beschrieben durchgeführt. Die Konzentration der extrahierten DNA wurde mit einem Qubit® Fluorometer unter Verwendung des Broad Range Kits wie im Handbuch beschrieben ermittelt. Die Präparation der DNA zur Sequenzierung erfolgte nach einem angepasstem Protokoll des Nextera XT DNA Library Preparation Kits von Illumina® [118]. Als Ausgangsmaterial wurden 0,5 ng/μl eingesetzt. Die Tagmentation wurde mit einem verringerten Volumen des TDE1 Enzyms von 0,0083 μl durchgeführt, der resultierende Mastermix hatte ein Volumen von 1,33 μl pro Probe. Die Konzentration der resultierenden DNA Bibliothek wurde durch ein Qubit® Fluorometer unter Verwendung des HS Assay Kits wie im Handbuch beschrieben durchgeführt. Zur Sequenzierung wurden nur DNA Bibliotheken mit einer Konzentration von mindestens 1 ng/μl eingesetzt.
Für ein zu sequenzierendes *C. difficile* Genom wurde eine Größe von 5 Megabyte (MB) angenommen und eine 70fache Abdeckung angestrebt, um bei einer zu erwartenden Ausgabe von zum Beispiel 35.000 MB (NextSeq Gerät) 100 DNA-Bibliotheken auf einem Sequenzierlauf sequenzieren zu können. Die Konzentration des resultierenden Pools wurde durch ein Qubit® Fluorometer unter Verwendung des HS Assay Kits wie im Handbuch beschrieben durchgeführt und anschließend auf eine Konzentration von 4mM gebracht. Die meisten DNA Bibliotheken Pools wurden auf dem Illumina® NextSeq Gerät mit V2 Chemie sequenziert. Vereinzelt wurden Sequenzierungen mit dem Illumina® MiSeq Gerät unter Verwendung von V3 Chemie durchgeführt. Auf beiden Geräten wurde die Sequenzierung im *Paired-end*-Modus mit 150 Zyklen pro Read (2x150) durchgeführt.

2.2 Prozessierung der Sequenzdaten

2.2.1 Qualitätsprüfung und Auswahl der Sequenzdaten

Vor der Sequenzrekonstruktion wurde die Qualität der generierten Illumina-*Reads* evaluiert. Dafür wurde die durchschnittlichen Abdeckung der *Reads* pro Base abgeschätzt, wofür das von Dr. Matthias Steglich (Leibniz-Institut DSMZ, Braunschweig) implementierte Skript *fastq-coverage.sh* verwendet [79] wurde. Um eine zuverlässige Sequenzrekonstruktion zu gewährleisten, wurde eine Abdeckung von 50 *Reads* pro Base angestrebt. Für Sequenzen, für die diese Abdeckung nicht im ersten Sequenzierlauf erreicht wurde, wurde die DNA Bibliothek erneut sequenziert. Anschließend wurden die Illumina-*Reads* aus zwei Sequenzierläufen konkateniert, um so die Abdeckung zu steigern. Genome, für die auch nach erneuter Sequenzierung keine ausreichende Abdeckung erreicht wurde, wurden von den Analysen ausgeschlossen.
Ein weiteres Kriterium war der in EnteroBase implementierte Qualitätscheck der resultierenden Assemblierungen nach Hochladen der Illumina-*Reads* auf die Plattform (Tabelle 2.1). Wurden die Kriterien hier nicht vollends erfüllt, wurde das zugehörige Genom für weitere Analysen ausgeschlossen.

Tabelle 2.1: Qualitätsparameter für die *C. difficile* Assemblierungen in EnteroBase. Mbp: Megabasenpaare; Kbp: Kilobasenpaare; N50: Anzahl an *Contigs* mit einer Länge, die über 50 % der gesamten Genomsequenz liegt; N: nicht eindeutig bestimmbare Base

Metriken	Kriterien
Anzahl der Basen	3.6Mbp - 4.8 Mbp
N50	>20Kbp
Anzahl der Contigs	<600
Anteil an N's	<5 %
Zuordnung der *Contigs* zu *C. difficile*	>65 %

Um die cgMLST-Analyse mit der SNP-Analyse zu vergleichen wurden neben den Datensätzen, die speziell für diese Arbeit sequenziert wurden, auch publizierte Daten analysiert. Auch diese wurden einer Qualitätskontrolle unterzogen. Im Folgenden werden die resultierenden Datensätze dargestellt.

Datensatz von Patienten mit wiederkehrender CDI

Von den 183 gewonnenen Isolaten wurden 176 (96 %) erfolgreich sequenziert und aufgrund der ausreichenden Abdeckung der *Reads* pro Base für weitere Analysen verwendet. Die FASTQ-Dateien wurden im Europäischen Nukleotid Archiv unter der *Study Accession number* PRJEB33768 hinterlegt. Eine Liste der Isolate befindet sich in Anhang A Tabelle A.1, Tabelle 2.2 zeigt die Verteilung der Isolate über die Patienten und Episoden.

Tabelle 2.2: Anzahl der Isolate isoliert aus vier Patienten mit wiederkehrender CDI. Die Wochenangabe bezieht sich auf die Zeitspanne zwischen der Diagnose während des ersten und zweiten Aufenthaltes des Patienten im Krankenhaus.

Patient	Δt (Wochen)	Anzahl der Isolate, Episode 1/2
D	21	15/14
E	11,4	28/36
F	16	32/35
G	21,9	12/4

Datensatz aus einem Netzwerk von Krankenhäusern

Für diesen Datensatz wurden 291 der 309 isolierten Isolate mit ausreichender Abdeckung sequenziert (94 %). Zwei dieser Isolate stammten vom gleichen Patienten und wurden aufgrund von Unterschieden in ihrer Morphologie und der dadurch entstandenen Vermutung, dass eine Infektion durch zwei *C. difficile* Stämme vorliegt, beide sequenziert. Die Analysen zeigten jedoch, dass die Isolate identisch waren. Aufgrund dessen wurde eines dieser Isolate von der Analyse ausgeschlossen, sodass für diesen Datensatz Genome von 290 Isolaten analysiert wurden. Die zugehörigen FASTQ-Dateien wurden unter der *Study Accession number* PRJEB33779 hinterlegt.

Tabelle 2.3: Anzahl der Isolate von Patienten mit CDI, die in einem der sechs beprobten Krankenhäuser hospitalisiert waren. Die Isolate wurden über einen Zeitraum von drei Perioden gesammelt.

Kran-kenhaus \ Proben-nahme	02.09.2013 – 11.7.2014	22.07.2014 – 15.04.2015	03.09.2015 – 06.04.2016
1	16	18	24
2	20	19	20
3	18	18	25
4	0	0	20
5	19	21	21
6	18	13	0

Eine Liste der Isolate mit zugehörigem Krankenhaus, Station und Probennahme, sowie weitere Stationen beziehungsweise Krankenhäuser in denen sich der jeweilige Patient vor oder nach der Diagnose befand, sind in Anhang A Tabelle A.2 zu finden. Tabelle 2.3 gibt einen Überblick über die Verteilung der Isolate über die Krankenhäuser und die Perioden.

SOARiAL Datensatz

Von den insgesamt 209 isolierten *C. difficile* Isolaten wurden 191 (91 %) erfolgreich sequenziert und die Genome durch EnteroBase analysiert. Tabelle 2.4 zeigt die Verteilung der Isolate auf die Proben. Eine

ausführliche Liste der Isolate mit Probenbezeichnung befindet sich in Anhang A Tabelle A.5.

Tabelle 2.4: Anzahl der untersuchten Isolate aus den verschiedenen Proben des SOARiAL Projekts. Bis auf das Staubisolat wurden die Isolate aus Proben des Feldversuches isoliert. Das Staubisolat wurde aus einer Probe des Windkanalversuchs isoliert.

Probe	Anzahl an Isolaten
Dünger	72
Sammelkotprobe	3
Boden	5
gedüngter Boden	21
gedüngter Boden (2 Wochen)	11
gedüngter Boden (4 Wochen)	9
gedüngter Boden (7 Wochen)	11
gedüngter Boden (10 Wochen)	21
gedüngter Boden (14 Wochen)	20
gedüngter Boden (19 Wochen)	17
Staub	1

Isolate mit PCR Ribotypen Information

Von den insgesamt 670 untersuchten Isolaten, die für die Studie von Zaiß et al. in 84 Krankenhäusern in ganz Deutschland gesammelt und anschließend ribotypisiert wurden, wurden 71 im Rahmen dieser Arbeit sequenziert [119]. Die Hinterlegung der FASTQ-Dateien erfolgte unter der *Study Accession number* PRJEB33868.
Um einen Einblick in die Populationsstruktur von *C. difficile* in Schweinen zu erhalten, wurden 184 Genome von Isolaten, die im Zuge der Arbeit von Schneeberg et al. von rektalen Abstrichen von Schweinen von 15 Bauernhöfen gesammelt und ribotypisiert wurden, sequenziert [86]. Die FASTQ-Dateien wurden unter der *Study Accession number* PRJEB33780 hinterlegt.
Des Weiteren wurden 108 Genome von Isolaten, die während einer unpublizierten Studie aus Stuhlproben aus Altenheimen in Deutschland isoliert wurden, sequenziert. Auch für diese Isolate wurde eine PCR Ribotypisierung durchgeführt. Die FASTQ-Dateien wurden unter der *Study Accession number* PRJEB33866 hinterlegt.
Die resultierenden Illumina-*Reads* und die zugehörigen Metadaten dieser Datensätze wurden der *C. difficile* DatenBank in EnteroBase hinzugefügt und standen somit für vergleichende Analysen zur Verfügung. Detaillierte Informationen zu den Isolaten befinden sich in Anhang A Tabelle A.4.

Vier publizierte Ausbrüche

Sowohl für die Korrelation der aus der cgMLST- und SNP-Analyse resultierenden genomischen Distanzen als auch für die Detektion von Ausbrüchen durch HC2 Cluster wurden vier publizierte Ausbrüche analysiert. Zwei dieser Ausbrüche sind Bestandteil der Publikation von García-Fernández et al. [55]. Insgesamt wurden in dieser Publikation 26 Isolate als PCR Ribotyp 027 typisiert, wovon nach der Qualitätskontrolle 22 Genome analysiert wurden. Als PCR Ribotyp 106/500 wurden 61 Isolate typisiert, wovon 34 einer von sieben detektierten Transmissionsketten zugehörig waren. Für die Analyse wurde die längste Transmissionskette mit 20 Isolaten ausgewählt. Die neun in der Publikation von Berger et al. aufgeführten Genome, dessen Isolate während eines Ausbruchs des PCR Ribotypen 018 isoliert wurden, entsprachen den Qualitäten der cgMLST- und SNP-Analyse [120]. Dies galt auch für die 13 Genome der RT027 Isolate in der Publikation von Jia et al. [112].

Oxfordshire Datensatz

Für den Vergleich der cgMLST- und SNP-Analyse wurde die Publikation, in der ein möglicher Zusammenhang von zwei *C. difficile* Isolaten durch eine Transmission mit einer genomischen Distanz von ≤2 SNPs definiert wurde, reanalysiert. Der in dieser Publikation untersuchte Datensatz umfasste 1.223 Genome [72]. Dabei wurden 266 Isolate in der so genannten „Einlaufperiode“ (01.09.2007–31.03.2008)und 957 Isolate in der „Testperiode“ (01.04.2008–31.03.2011) isoliert. Die Einlaufperiode startete 30 Wochen vor der eigentlichen Testperiode, um mögliche Quellen für Infektionen in der Testperiode zu erfassen. Dabei wurde angenommen, dass ein erkrankter Patient eine Woche vor und acht Wochen nach der Diagnose infektiös war. Zusätzlich wurde eine Inkubationszeit von 0-12 Wochen angenommen. Um die genaue Anzahl an Fällen zu zählen, die aufgrund einer Transmission durch einen vorher hospitalisierten Patienten erkrankten, wurden nur die Isolate gezählt, die in einem paarweisen Vergleich von ≤2 SNPs das spätere Datum hatten, somit also als Akzeptor der Infektion galten. Hierbei ist darauf zu achten, dass Donor und Akzeptor nicht beide während der Einlaufperiode isoliert wurden.

Für diesen Datensatz konnten 1.158 Genome in der *C. difficile* Datenbank gefunden werden. Die Illumina-*Reads* der fehlenden Genome wurden aus dem Europäischen Nukleotid Archiv runtergeladen und manuell der Datenbank hinzugefügt. Dabei stellte sich heraus, dass die resultierenden Assemblierungen nicht den Qualitätsstandards der cgMLST-Analyse entsprachen. Somit wurde die Reanalyse mit 1.158 Genomen durchgeführt; 242 in der Einlaufperiode und 916 in der Testperiode. Die SNP Distanzen wurden von Dr. David Eyre (Universität Oxford) zur Verfügung gestellt.

2.2.2 Bioinformatische Analysen

Die Illumina-*Reads* der ausgewählten Datensätze wurden zum einen mit der cgMLST-Analyse analysiert, die von den in EnteroBase implementierten bioinformatischen Werkzeugen Gebrauch macht, zum anderen wurden die Illumina-*Reads* mit der *Mapping*-basierten SNP-Analyse prozessiert. Da die SNP-Analyse rechenintensiv ist und die Analyse von umfangreichen Datensätzen viele Ressourcen und Zeit beansprucht, wurde für solche die auf Assemblierungen basierende SNP-EToKi-Analyse angewendet. Ausgewählte Datensätze wurden zusätzlich mit der Software SeqSphere$^+$ analysiert, um das dort verwendete cgMLST Schema mit dem in EnteroBase vergleichen zu können. Eine weitere Alternative zur Bestimmung von genomischen Verwandtschaften bietet die Software PopPUNK, die zur Bestimmung der Unterschiede im Gengehalt des akzessorischen Genoms herangezogen wurde. Eine Übersicht der angewendeten Analysen und deren Ergebnis bietet Abbildung 2.1.

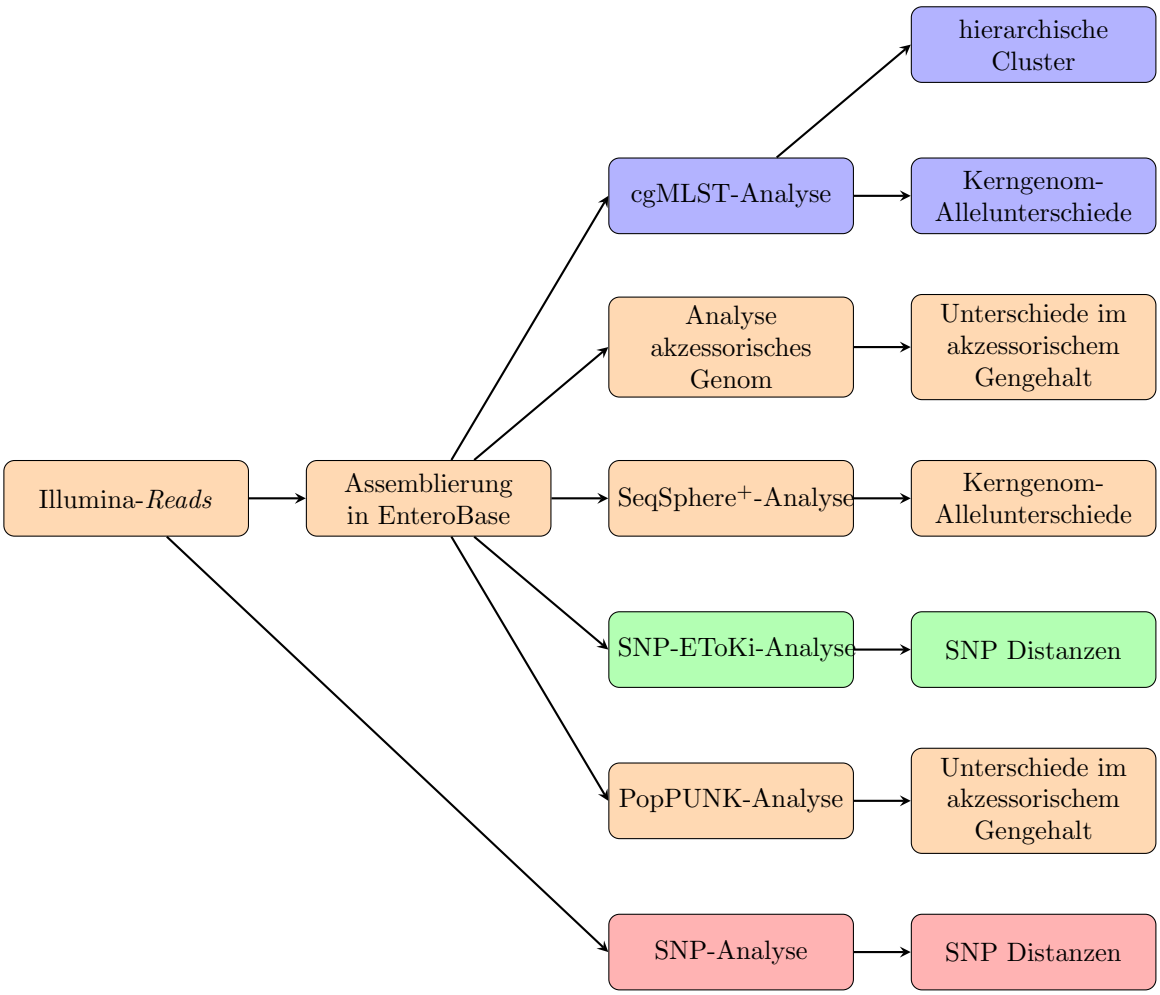

Abbildung 2.1: Übersicht der durchgeführten Analysen für die Illumina-*Reads* ausgewählter Datensätze und deren zugehöriges Ergebnis. Die eingefärbten Analysewege werden in den Abbildungen 2.2, 2.3 und 2.4 detaillierter dargestellt.

cgMLST-Analyse

Eine detaillierte Übersicht aller in EnteroBase implementierten bioinformatischen Werkzeuge und Pipelines befindet sich in der Publikation von Zhou et al. [68]. Folgend werden nur die für die Analyse relevanten Schritte erläutert (Abbildung 2.2).

Zu Beginn der cgMLST-Analyse wurden die Illumina-*Reads* auf die EnteroBase Plattform hochgeladen. Anschließend wurden die *Reads* mit der EBAssembly Pipeline assembliert. Die Pipeline besteht aus 2 Hauptschritten: Zunächst werden die *Reads* getrimmt und, falls nötig, auf eine Abdeckung von maximal 100 reduziert („EToKi prepare“). Die folgende De-Novo-Assemblierung wird mit SPAdes (Version 3.10 [121]) durchgeführt. Um mögliche Fehler in diesem Prozess zu vermeiden, werden die der Assemblierung zugrunde liegenden *Reads* an diese aligniert (BWA, [122]) und mit Pilon [123] auf Diskontinuitäten, Insertionen/Deletionen und Unterschiede in einzelnen Basen untersucht. Pilon korrigiert beziehungsweise reassembliert die detektierten Stellen anschließend. Nachdem aus der resultierenden Assemblierung *Contigs* mit einer Abeckung <30 entfernt und die Anzahl der nicht eindeutig bestimmbaren Basen („N“) abgeschätzt wird, erfolgt die Evaluierung der wahrscheinlichsten taxonomischen Quelle der Sequenzen (Kraken [124]). Diese Schritte sind Teil des Programms „EToKi assemble“.

Die Assemblierungen, die den Kriterien in Tabelle 2.1 entsprechen, werden durch die MLSType Pipeline weiter prozessiert [68]. Durch diese erfolgt für jede Assemblierung die Bestimmung der Allelprofile für jedes der in EnteroBase verfügbaren MLST Schemata. EnteroBase pflegt eine Datenbank von so genannten „exemplarische Allelen“, die aus einer Allelsequenz für jedes Gen in dem Ganzgenom-MLST Shema

(wgMLST) besteht. Diese exemplarischen Allele werden mit BLASTn [125] gegen die Assemblierung aligniert, sodass beide Enden des exemplarischen Allels abgedeckt werden. Dieser so bestimmten Sequenz wird anschließend eine Allelnummer zugeordnet und dem dadurch entstehendem MLST Allelprofil ein Sequenztyp (ST).

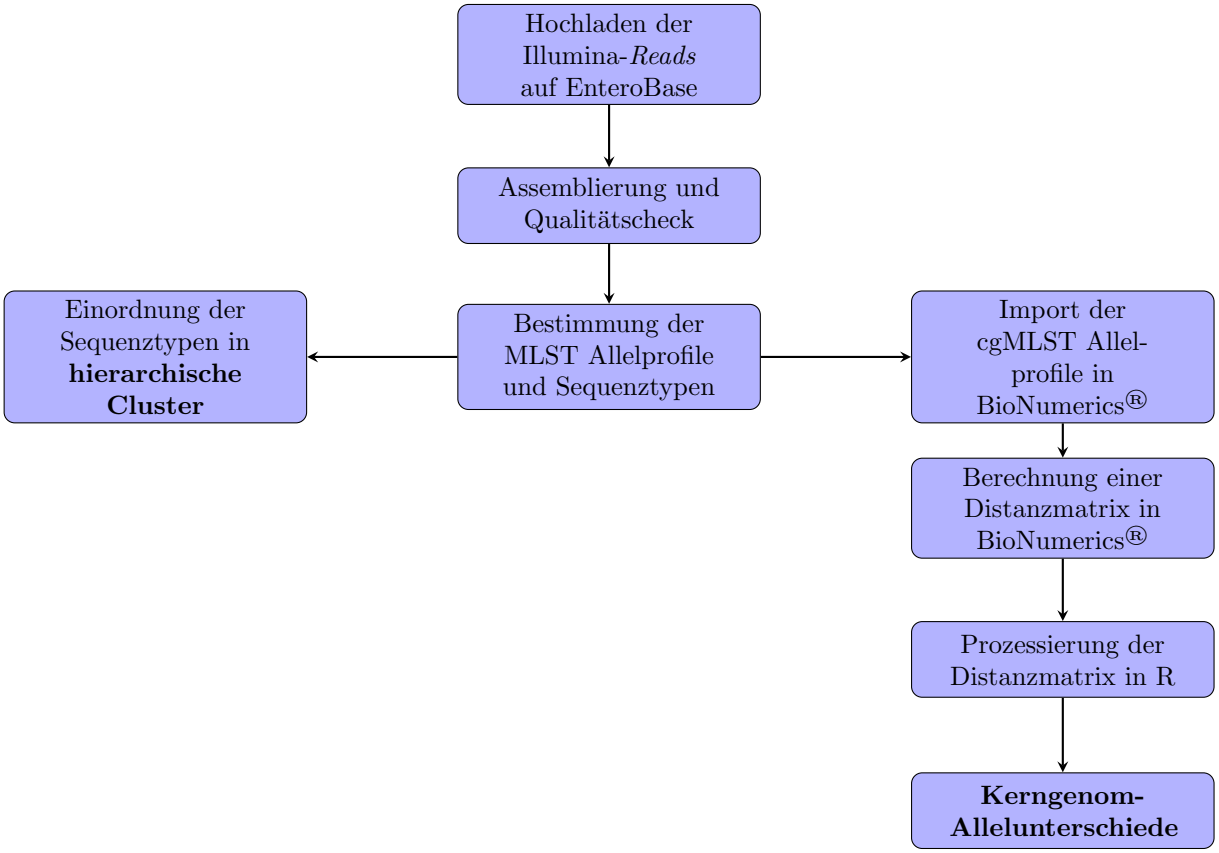

Abbildung 2.2: Ablauf der cgMLST-Analyse der qualitätskontrollierten Illumina-*Reads*. Nachdem die *Reads* auf die EnteroBase-Plattform hochgeladen wurden, wurden diese automatisch duch eine Pipeline assembliert und anschließend die Qualität der resultierenden Assemblierungen überprüft. Für Assemblierungen mit ausreichender Qualität wurden folgend Allelprofile für alle in EnteroBase verfügbaren MLST Schemata bestimmt. Auf Basis der durch die cgMLST Allelprofile bestimmten Kerngenom-Sequenztypen wurden den Assemblierungen hierarchische Cluster auf insgesamt 13 Ebenen zugeordnet (HC0, 2, 5, 10, 20, 50, 100, 150, 200, 500, 950, 2000 und HC2500). Des Weiteren wurden die cgMLST Allelprofile zur Berechnung der paarweisen Distanzen in die Software Bionumerics® importiert. Die berechnete Distanzmatrix wurde dann in R in paarweise Vergleiche der Kerngenom-Allelunterschiede prozessiert.

Befindet sich die Allelsequenz noch nicht in der Datenbank und verfügt über eine ausreichende Qualität, so wird diese in die Datenbank mit aufgenommen und einer neuen Allelnummer zugeordnet. Liegt eine Deletion an einer alignierten Stelle der Assemblierung vor, so wird dies mit einer Allelnummer 0 gekennzeichnet. Wenn die Assemblierungssequenz aufgrund von Fragmentierung, Duplikation oder schlechter Sequenzqualität keiner Allelnummer zugeordnet werden kann, wird dies mit „-“ gekennzeichnet und im Folgenden als unbestimmtes Allel behandelt. Das in dieser Arbeit schwerpunktmäßig genutzte cgMLST Schema besteht aus einer Teilmenge von Loci aus dem wgMLST Schema, die in 95 % der Referenzgenome, die zur Erstellung des wgMLST Schematas verwendet wurden, vorkamen.
Anschließend wurden die cgMLST Allelprofile in die Software BioNumerics® importiert, um eine Distanzmatrix basierend auf den cgMLST Allelprofilen der Einträge zu berechnen. Da es sich bei den Allelprofilen um kategoriale Daten handelt und die Unterschiede zwischen zwei Profilen berechnet werden sollte, wurde der kategorische Koeffizient für Unterschiede zur Berechnung der Distanzen angewendet. Dieser berechnet die Anzahl der Allele, in denen sich die beiden Allelprofile unterscheiden, wobei nicht

bestimmte Allele paarweise von dem Vergleich ausgeschlossen werden. BioNumerics® ermöglicht allerdings nur Berechnungen bis zu einer Distanz von 200 Allelunterschieden. Um größere Unterschiede zwischen den Allelprofilen trotzdem erfassen zu können wurde bei der Berechnung ein Faktor von 100 angewendet. Die resultierende Distanzmatrix wurde dann für weitere Analysen in R importiert und mit der Funktion *dm_to_dlist.R*[126] in paarweise Distanzen umgewandelt.

An- und Abwesenheitsunterschiede im akzessorischem Genom

Um Genome auch in ihrem akzessorischen Gengehalt vergleichen zu können, wurden die wgMLST Allelprofile um die Loci, die im cgMLST Allelprofil vorkommen, gekürzt und anschließend mit der Funktion *access_to_binary.R*[127] binär kodiert. Wenn ein Genom an einer Locusstelle eine Allelsequenz aufweist, wurde eine 1 zugeordnet. Dabei wurden nicht bestimmte Allelsequenzen als 'NA' markiert, damit diese paarweise von dem Vergleich ausgeschlossen werden können. Wenn an der Locusstelle keine Allelsequenz detektiert wurde, wurde eine 0 zugeordnet. So konnten die Genome in ihrem akzessorischem Gengehalt verglichen werden. Die paarweisen Distanzen der binär kodierten akzessorischen Loci wurden mit der *dist*-Funktion des R Pakets *stats* (Version 3.6.1) bestimmt ([116]). Dabei wurde die Methode *binary* gewählt.

SeqSphere$^+$-Analyse

Für ausgewählte Datensätze wurden die in EnteroBase generierten Assemblierungen in die kommerzielle Software SeqSphere$^+$ geladen und dort genotypisiert. Die entstandenen cgMLST Allelprofile wurden analog zu den cgMLST Allelprofilen aus EnteroBase in BioNumerics® importiert und wie in 2.2.2 beschrieben weiter prozessiert.

Assemblierungsbasierte SNP-EToKi-Analyse

Die Berechnung der SNP Distanzen durch die SNP-EToKi-Analyse wurde für umfangreiche Datensätze durchgeführt, da hierfür die schon vorliegenden Assemblierungen aus EnteroBase verwendet werden konnten und die ersten beiden Schritte der SNP-Analyse damit umgangen wurden (siehe Abbildung 2.4). Im Folgenden werden die Module des „EnteroBase Tool Kit“ (EToKi) so beschrieben, wie sie zum Zeitpunkt der Analysen zur Verfügung standen und angewendet wurden (Abbildung 2.3; Stand vom 21.7.2019;[128]). Die Assemblierungen wurden über ein durch Kooperationsarbeit entstandenes Python-Skript [129] von der EnteroBase Plattform heruntergeladen, wodurch dieser Schritt automatisiert und zeiteffizienter durchgeführt werden konnte. Das Alignment der Assemblierungen gegen eine Referenzsequenz wurde mit Hilfe des *EToKi.py align* Kommandos erstellt. Folgend wurde mit dem Kommando *EToKi.py phylo* ein Maximum-Likelihood phylogenetischer Baum berechnet, wobei hier das Programm RAxML (Version 8 [130]) verwendet wurde. Die Detektion von Rekombinationen (*EToKi.py RecHMM*) erfolgte mit dem Programm RecHMM [131]. Anschließend wurden mit dem Kommando *EToKi.py RecFilter* SNPs, die laut der RecHMM-Analyse durch eine Rekombination hervorgerufen wurden, in dem Alignment maskiert und so wie eine nicht eindeutig bestimmbare Base behandelt. Diese nicht eindeutig bestimmbaren Basen wurden durch die Anwendung der *snp-dists*-Funktion [132] paarweise bei der Erstellung der SNP-Distanzmatrix ausgeschlossen. Die entstehende SNP-Distanzmatrix wurde in R weiter prozessiert.

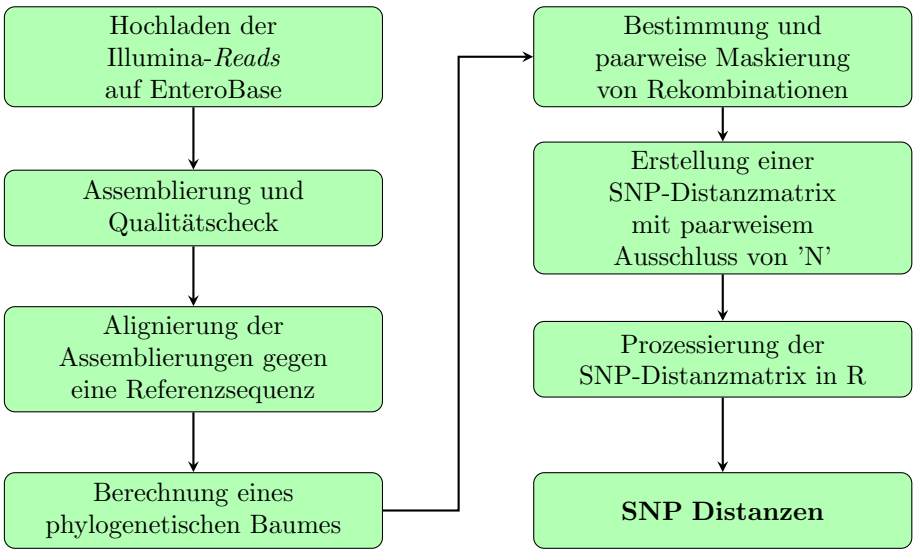

Abbildung 2.3: SNP-EToKi-Analyse der qualitätskontrollierten Illumina-*Reads*. Nachdem die *Reads* auf die EnteroBase Plattform hochgeladen und automatisch assembliert wurden, folgt eine Qualitätsüberprüfung der Assemblierungen. Die *Contigs* der Assemblierungen mit ausreichender Qualität wurden gegen eine Referenzsequenz aligniert. Mit dem entstehendem Alignment wurde ein Maximum-Likelihood phylogenetischer Baum berechnet. Anschließend wurden mit Hilfe des Alignments und des Baumes die Rekombinationen bestimmt. Im Folgendem wurden die SNPs, die als Folge einer Rekombination entstanden, in dem Alignment maskiert. Auf Basis dieses Alignments wurden eine SNP-Distanzmatrix berechnet. Hierbei wurden nicht bestimmte Basen sowie die maskierten Rekombinationen paarweise ausgeschlossen. Die entstandene SNP-Distanzmatrix wude in R weiter prozessiert.

Read-Mapping basierte SNP-Analyse

Die SNP-Analyse wurde hauptsächlich mit Hilfe von Shell-Skripten von Dr. Matthias Steglich, die die Handhabung der verwendeten bioinformatischen Werkzeuge und die Bearbeitung des Alignments automatisierten, durchgeführt. Ein Überblick der einzelnen Schritte ist in Abbildung 2.4 dargestellt. Das *Read-Mapping* gegen eine Referenzsequenz erfolgte mit Hilfe des Burrows-Wheeler Aligners (Version 0.7.12 [133]) unter Verwendung des BWA-MEM-Algorithmus. Dieser Algorithmus lässt das Aufsplitten von *Reads*, die an zwei verschiedenen Stellen der Referenzsequenz eine Übereinstimmung zeigen, zu. Es wurden die Standardeinstellungen verwendet, mit Ausnahme der Anpassung der *seed*-Länge für ausgewählte Analysen auf 30 Nukleotide (siehe Kapitel 3.5.1). Die entstandenen Dateien wurden mit SAMtools (Version 0.1.19 [134]) in binäre *Sequence-Alignment-Map*-Format Dateien (BAM-Dateien) konvertiert, sodass anschließend durch das Werkzeug VarScan2 (Version 2.3 [135]) der Konsensus an jeder Position der Referenzsequenz ermittelt werden konnte und in einer Konsensussequenz im FASTA-Format resultierte. Die FASTA-Dateien der Genome, die miteinander verglichen werden sollten, wurden zusammen mit der Referenzsequenz zu einem Alignment zusammengefügt. Für die Rekonstruktion des Stammbaumes unter Verwendung der Maximum-Likelihood Methode wurde das Programm RAxML (Version 8.2.9 [130]) mit Hilfe der Shell-Skripte von Dr. Markus Göker (Leibniz-Institut DSMZ, Braunschweig) angewendet. Nach der Überführung des Alignments in das *extended PYHLIP*-Format wurde die Berechnung des Stammbaumes durch RAxML unter Verwendung des *Bootstopp*-Kriteriums durchgeführt. RAxML erkannte so automatisch wieviele *Bootstrap*-Replikate nötig waren um stabile Unterstützungswerte für den besten Baum zu erreichen. Auf Basis des resultierenden Stammbaumes wurden anschließend mit Hilfe der Software ClonalFrameML (Version 1.11 [74]) Rekombinationen detektiert und diese anhand der Koordinaten spaltenweise aus dem Alignment entfernt.

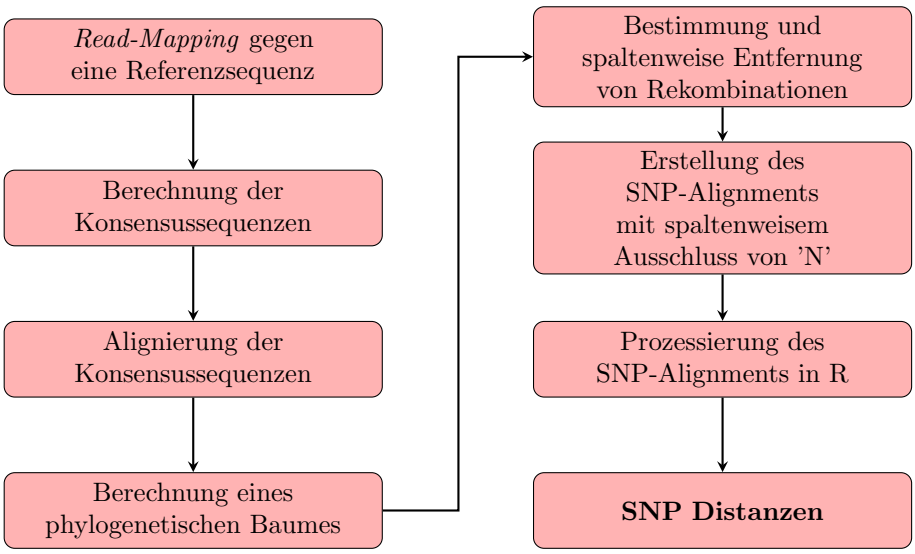

Abbildung 2.4: Ablauf der SNP-Analyse der qualitätskontrollierten Illumina-*Reads*. Nachdem die *Reads* gegen eine Referenzsequenz aligniert wurden, wurde die Konsensussequenz ermittelt. Anschließend wurden alle zu analysierenden Konsensussequenzen mit der für den *Mapping*-Prozess verwendeten Referenzsequenz in ein Alignment zusammen gefasst und ein Maximum-Likelihood phylogenetischer Baum berechnet. Mithilfe des berechneten Baumes und des Alignments wurden folgend die Koordinaten von Rekombinationen ermittelt. Die bestimmten Koordinaten wurden spaltenweise aus dem Alignment entfernt. Anschließend wurde das Alignment auf die Mutationen enthaltenen Stellen gekürzt und die paarweisen SNPs in R berechnet.

Das Alignment wurde weiter auf die Mutationen enthaltenen Stellen gekürzt, wobei Spalten, die nicht eindeutig bestimmbare Basen („N“) beinhalteten, ausgeschlossen wurden. Für das resultierende SNP-Alignment wurden in R mit Hilfe der Funktion *alignment_to_snpslist.R*[136], die die *stringDist*-Funktion des *Biostrings*-Pakets (Version 2.54.0 [137]; „hamming“-Methode) nutzt, die paarweisen SNP-Distanzen bestimmt.

Verwendete Referenzsequenzen

Als Referenzsequenz sollte ein geschlossenes *C. difficile*-Genom verwendet werden, dass nah verwandt mit dem zu untersuchenden Datensatz ist. So wurde in anderen Studien das Genom des Stammes M120 für RT078 Datensätze [56], [138] , CD196 und R20291 für RT027 Datensätze [72], [139], [140] und CD630 für diverse Datensätze [71], [72] für den *Read-Mapping*-Prozess verwendet (Tabelle 2.5).

Tabelle 2.5: Referenzsequenzen die für die Alignierung der Illumina-*Reads* in der SNP-Analyse beziehungsweise der *Contigs* der Assemblierungen in der SNP-EToKi-Analyse verwendet wurden.

Referenz	Accession-Nummer	PCR Ribotyp
R20291	FN545816	RT027
CD630	CP010905.2	RT012
CD196	NC_013315.1	RT027
M120	NC_017174.1	RT078

In der vorliegenden Arbeit wurde das Genom des R20291 Stammes als Referenzsequenz für die SNP-Analyse der Genome der Isolate aus dem Datensatz der vier publizierten Ausbrüche und dem Datensatz der wiederkehrenden CDI ausgewählt. Für die Genome der Isolate von Patienten mit wiederkehrender CDI lag die Information für den PCR Ribotypen nicht vor, sodass keine nah verwandte Sequenz gewählt werden konnte. Zwei der vier publizierten Ausbrüche wurden durch Ribotyp 027 Isolate verursacht und waren somit nah verwandt mit der ausgewählten Referenzsequenz. Für den PCR Ribotyp 018 Ausbruch lag kein geschlossenes Genom mit gleichem PCR Ribotypen vor. Die Illumina-*Reads* der Isolate des RT106

Ausbruchs wurden zusätzlich gegen die M120 Referenzsequenz aligniert, resultierten aber in gleichen SNP Distanzen, sodass für die in dieser Arbeit präsentierten Ergebnisse die SNP Distanzen der SNP-Analyse mit der R20291 Referenzsequenz verwendet wurden. Für die SNP-Analyse des hoch diversen Datensatzes der 816 zur Verfügung stehenden Illumina-*Reads* der 1.004 Isolate mit einer genomischen Verwandtschaft von $\leq$2 Kerngenom-Allelunterschieden zu einem Isolat aus einem anderen Land (Kapitel 3.4.4) wurde die R20291 Referenzsequenz verwendet.

Variationen der SNP-Analyse

Um die Auswirkungen der einzelnen Schritte in der SNP-Analyse auf die resultierende Anzahl der genomischen Distanzen zwischen Isolaten nachvollziehen zu können, wurden diese variiert. Zum einen wurde die SNP-Analyse sowohl mit Einbezug der rekombinativen Sequenzabschnitte als auch ohne Rekombinationen durchgeführt. Der Ausschluss der Rekombinationen, die mit ClonalFrameML oder RecHMM detektiert wurden, erfolgte dabei entweder wie für die SNP-Analyse beschrieben spaltenweise, oder paarweise (siehe Abbildungen 2.4 und 2.3). Eine weitere Möglichkeit bat ein von Dr. David Eyre zur Verfügung gestelltes Python-Skript. Dieses rechnete die Astlängen des phylogenetischen Baumes, der aus der ClonalFrameML-Analyse resultierte und somit auf Rekombinationen korrigiert war, in SNPs um. Der daraus resultierende Baum wurde mit der *cophenetic* Funktion des R Pakets *stats* (Version 3.6.1) [116] in eine Distanzmatrix konvertiert und weiter prozessiert. Des Weiteren wurde der Schritt der SNP Detektion variiert. Das Alignment wurde entweder um Spalten, welche nicht eindeutig bestimmbare Basen enthielten, gekürzt (Abbildung 2.4) oder wie in Abbildung 2.3 beschrieben durch paarweisen Ausschluss in eine Distanzmatrix umgerechnet. Die resultierende Distanzmatrix wurde mit der Funktion *alignment_to_snpslist.R*[136] in paarweise Vergleiche transferiert.

Verwendung der Software PopPUNK

Neben der Bestimmung der auf wgMLST-Allelprofilen basierenden Unterschiede des Gengehalts im akzessorischem Genom wurden diese auch mit der Software PopPUNK berechnet (*POPulation Partitioning Using Nucleotide K-mers* [141]). PopPUNK erstellt auf variablen *k-mer*-Längen basierend stark reduzierte Darstellungen der Assemblierungen, so genannte *sketches*. Für langsam evolvierende Bakterien wie *C. difficile* wird eine *sketch*-Größe von 10^5 empfohlen [141]. Das bedeutet, dass jede Assemblierung in 10^5 Positionen eingeteilt wird. Diese Positionen nehmen *hash*-Nummern ein, die, ähnlich wie in der cgMLST-Analyse, einer individuellen *k-mer*-Sequenz zugeordnet werden. PopPUNK erstellt diese *sketches* für unterschiedliche *k-mer*-Längen, wobei die geringste *k-mer*-Länge k_{min} so gewählt wird, dass Repetitionen vermieden werden. Für *C. difficile*-Assemblierungen wurde eine k_{min} von 17 gewählt, die längste *k-mer*-Sequenz war mit 29 Basen voreingestellt. Es wurden für jede Assemblierung vier *sketches* erstellt (*k-mer*-Längen von 17, 21, 25 und 29 Basen). Die den *hash*-Nummern zugrunde liegenden k_{min}-Sequenzen werden für jeden untersuchten Datensatz in Kern- und akzessorisches Genom eingeteilt. Hierbei erfolgt die Deklaration einer Stelle in der *sketch* als akzessorisches Genom, wenn eine der zu vergleichenden Sequenzen sich in jeder Base von den anderen Sequenzen in dieser Position unterscheidet oder eine *sketch* an der Position keine zugeordnete Sequenz vorweist (Deletion). Anschließend berechnet PopPUNK zwei Wahrscheinlichkeiten, deren Produkt die Wahrscheinlichkeit eines übereinstimmenden *k-mers* zwischen zwei Sequenzen ergibt. Zum einen wird die Wahrscheinlichkeit, dass eine Kerngenom-Position eine Sequenz enthält, die keine Mutationen zu anderen Sequenzen zeigt, berechnet. Zum anderen wird die Wahrscheinlichkeit, dass sich ein Paar in der An-/Abwesenheit von *k-mer*-Sequenzen, die dem akzessorischem Genom zugeordnet wurden, unterscheiden, bestimmt. Die so beschriebene akzessorische Divergenz ist als Jaccard-Distanz definiert [141]. Die Abschätzung dieser Wahrscheinlichkeiten erfolgt durch den *MinHash*-Algorithmus, der Überschneidungen zwischen den Kerngenom-Distanzen, sowie den Jaccard-Distanzen des akzessorischen Genoms der vier *sketches* detektiert und so die am wahrscheinlichsten vorkommenden genomischen Distanzen berechnet.

In dieser Arbeit wurden die mittels der PopPUNK-Analyse berechneten Jaccard-Distanzen im akzessorischen Genom weiter prozessiert. PopPUNK erzeugte als eines der Endergebnisse eine Liste an paarweisen Vergleichen der untersuchten Assemblierungen mit den zugehörigen Distanzen im Kern- und akzessorischem Genom.

Darstellung der genomischen Beziehungen

Die genomischen Beziehungen zwischen Isolaten von ausgewählten Datensätzen sowie der Überblick über die Populationsstruktur von *C. difficile* und der Vergleich zwischen der PCR Ribotypisierung und HC150 Clusterung wurden durch das in EnteroBase implementierte Programm GrapeTree dargestellt [115]. Die Nachbearbeitung erfolgte in Inkscape (Version 0.92.3-1). Zur Darstellung wurden die in GrapeTree implementierten Algorithmen für Minimum Spanning (MST V2) und für Neighbor-Joining Bäume (RapidNJ) angewendet.

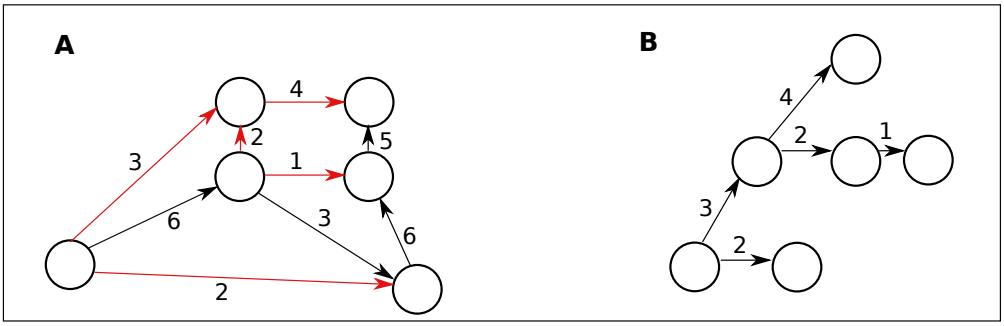

Abbildung 2.5: Vereinfachte Darstellung der Erstellung eines Minimum-Spanning Baumes. Zunächst wird ein gerichteter Arboreszenz Baum erstellt, in dem alle Knoten durch einen Pfad von der Wurzel aus erreichbar sind und die asymmetrischen Zweige die genomische Distanz basierend auf Kerngenom-Allelunterschieden wiedergeben **(A)**. Es werden die Zweige ausgewählt, durch die alle Knoten mit der geringsten Endsumme (Addition aller genomischen Distanzen der ausgewählten Zweige) verknüpfbar sind. Anschließend erfolgt eine lokale Neuzuordnung, um störende Verzweigungen zu entfernen und die Astlängen proportional zur Genomdistanz darzustellen **(B)**.

Beide Algorithmen basieren auf dem „Minimum Evolution“ Kriterium und auf Distanzmatrizen der genomischen Unterschiede der Isolate in den untersuchten Datensätzen. Bei Minimum-Spanning Bäumen wird in einem vollständig gegabelten Baum nach der Verbindung aller Knoten mit der geringsten Endsumme, also der aufaddierten Gesamtdistanz an Kerngenom-Allelunterschieden, gesucht ([142]; Seiten 405-406). Der in GrapeTree verwendete *MST V2* Algorithmus ordnet anschließend die Knoten entsprechend der minimalsten Verbindung neu an und skaliert die Astlängen proportional zu den Kerngenom-Allelunterschieden (Abbildung 2.5; [115]). Der Neighbor-Joining Algorithmus wiederum sucht iterativ nach der geringsten Verbindung zwischen zwei Knoten, wobei nach jeder gefundenen Verbindung die Distanzen zwischen den verbleibenden Knoten erneut berechnet werden. Der in GrapeTree verwendete Algorithmus *RapidNJ* beschleunigt die Suche nach verbindbaren Knoten, indem Knoten, die nicht für die nächste Verbindung in Frage kommen, von der Suche ausgeschlossen werden [143].

2.3 Statistische Untersuchungen

2.3.1 Evaluierung des cgMLST Schemas

Die Korrelation zwischen der Alleldiversität und der Locuslänge wurde mit dem Spearman'schen Rangkorrelationskoeffizienten berechnet, da die Datenpunkte keiner Normalverteilung entsprachen. Der Wert wurde mit der *cor.test*-Funktion aus dem *stats* Paket in R berechnet (Version 3.6.3)[116]. Das Auflösungsvermögen des hierarchischen Clusters HC0 wurde durch Simpson's Diskriminierungsindex

bestimmt. Die Berechnung wurde mit Hilfe der *diversity*-Funktion des *vegan* Pakets in R durchgeführt (Version 2.5-6 [144]).

2.3.2 Die Populationsstruktur von *C. difficile*

Übereinstimmung HC150 Cluster und PCR Ribotypisierung

Die Kongruenz zwischen der standardmäßig zur Typisierung von *C. difficile* Isolaten verwendeten PCR Ribotypisierung und dem in EnteroBase verfügbaren und auf der cgMLST-Analyse basierendem hierarchischem Cluster HC150, wurde durch den Adjusted Rand Koeffizienten [145] abgeschätzt. Hierfür wurde das online verfügbare *Comparing Partitions* Werkzeug verwendet (`http://www.comparingpartitions.info/`).

Rarefaction Methode

Um abschätzen zu können, inwieweit die Datenbankeinträge in EnteroBase die Vielfalt der *C. difficile* Genome und deren unterschiedlichen Stufen der Populationsstruktur erfassten, wurde die Rarefaction Methode zur Berechnung der Erwartungswerte für die hierarchischen Cluster HC150, 950, 2000 und 2500 verwendet. Die Berechnung erfolgte in R mit dem *iNEXT*-Paket (Version 2.0.2 [146]). Dabei wurde eine Diversitätsordnung $q = 0$ und *datatype* = abundance gewählt, da die Vielfalt für jedes hierarchische Cluster individuell berechnet werden sollte. Es wurde eine Probengröße von *knots* = 400 eingestellt, sodass die Diversitätsschätzungen für 400 gleichmäßig verteilte Knoten zwischen 1 und 27.030 (doppelte Anzahl der verfügbaren Genome in EnteroBase (13.515)) erfolgte.

2.3.3 Quantitativer Vergleich der SNP- und cgMLST-Analyse

Der lineare Zusammenhang zwischen den Kerngenom-Allelunterschieden und den SNP Distanzen wurde durch die *lm*-Funktion aus dem R Paket *stats* (Version 3.6.1)[116] berechnet. Die SNPs wurden mit der SNP-Analyse unter der Verwendung der Referenzsequenz R20291 bestimmt. Die SNP Distanzen der Genome, die Teil des Oxfordshire Datensatzes waren, stellte Dr. David Eyre zur Verfügung.
Um die Wahrscheinlichkeit, dass bei einem gewissen Wert an Kerngenom-Allelunterschieden die entsprechende SNP Distanz $\leq$2 ist, zu berechnen, wurde die lineare Regression für den Oxfordshire Datensatz zu einem verallgemeinerten linearen Modell erweitert (*GLM, generalized linear model*). Dazu wurden die von Dr. David Eyre zur Verfügung gestellten SNP Distanzen als vorherzusagende Variabel und die durch die cgMLST-Analyse berechneten Kerngenom-Allelunterschiede als Prädikator eingesetzt. Die SNP Distanzen wurden in eine binäre Variable konvertiert (0 = $\geq$2 SNPs; 1 = $\leq$2 SNPs), sodass zur Ermöglichung einer linearen Modellierung die Anwendung einer logistischen Regression nötig war ([147], Seiten 593-609). Dafür wurde für die SNP Distanzen eine Binomialverteilung angenommen und als *Link-Funktion* die Logit-Funktion gewählt. Die Analyse wurde mit Hilfe eines angepassten R-Skripts von Dr. Jan Meier-Kolthoff durchgeführt [148].

2.3.4 Lokale und globale Epidemiologie

Die Assoziation zwischen den 23 HC2 Clustern, in die sich die Genome der Isolate, die aus einem regionalen Netzwerk von Krankenhäusern isoliert wurden, durch die cgMLST-Analyse einteilen ließen, und den Krankenhäusern beziehungsweise den Stationen konnte aufgrund der Datenlage nicht mit klassischen Chi-Quadrat Tests (χ^2) durchgeführt werden. Der klassische χ^2-Tests erforderte in mindestens 80 % der Zellen der Kontingenztafel Werte über fünf und der alternativ für eine geringe Anzahl an Beobachtungen entwickelte exakte Test nach Fisher wurde für 2x2-Kontingenztafeln entwickelt. Aufgrund dessen wurden die Kontingenztafeln (Tabellen 2.6 und C im Anhang) mit 1.000 simulierten Datensätzen verglichen, in denen die Anzahl der Isolate in den Krankenhäusern beziehungsweise Stationen innerhalb der HC2 Cluster permutiert wurde. Für jede Reihe der originalen sowie der simulierten Kontingenztafeln wurde die normalisierte Shannon

Entropie und der χ^2-Wert berechnet (R Paket *entropy* Version 1.2.1. [149]). Die Werte wurden anschließend mit einem nichtparametrischen, zweiseitigem Mann-Whitney-U-Test (R-Paket *stats* Version 3.5.0 [116]) auf signifikante Unterschiede verglichen.

Tabelle 2.6: Kontingenztafel für die Genome der Isolate, die in einem regionalen Netzwerk von Krankenhäusern isoliert wurden. 133 Isolate dieser Studie wurden durch die cgMLST-Analyse in 23 HC2 Cluster eingeteilt. Die Tabelle zeigt die Verteilung der Isolate pro HC2 Cluster über die beprobten Krankenhäuser an.

HC2 Cluster \ Krankenhaus	1	2	3	4	5	6
1	0	0	1	0	2	0
2	0	0	2	0	0	0
70	0	0	3	0	0	0
76	15	13	19	2	9	8
85	0	0	2	0	0	0
109	0	0	0	0	1	2
479	0	0	0	0	2	0
491	2	0	0	0	0	0
1127	0	1	0	1	6	0
1131	0	1	1	1	2	0
1206	1	3	1	0	0	1
1208	1	0	1	0	0	1
1210	0	1	0	0	0	1
1225	2	0	1	0	0	0
1232	0	0	0	0	1	1
1242	0	0	0	1	0	1
1243	1	0	0	0	1	1
1251	0	0	2	0	1	2
1267	0	2	1	0	0	0
4415	0	2	0	0	0	0
4431	0	0	0	2	0	0
4808	0	0	0	0	2	0
4823	0	2	0	0	0	0

Um evaluieren zu können, ob die Differenzen im akzessorischen Gengehalt zwischen epidemiologisch nicht zusammenhängenden Isolaten mit fast identischen Kerngenomen ein normales Verhalten widerspiegeln, wurden mit der *sample_n*-Funktion aus dem R-Paket *dpyr* (Version 0.8.5 [150]) drei mal 1.000 Einträge zufällig aus den 13.515 Einträgen ausgewählt. Für diese drei Datensätze wurden die An-/Abwesenheitsunterschiede im akzessorischem Genom sowohl mit den akzessorischen Loci als auch mit PopPUNK bestimmt. Der Vergleich der Verteilungen der An-/Abwesenheitsunterschiede im akzessorischem Genom und der Zeitspanne für ausgewählte Isolatepaare mit $\leq$2 Kerngenom-Allelunterschieden aus bestimmten Datensätzen erfolgte mit dem *multcomp*-Paket in R (Version 1.4-12 [151]). Hierfür wurde zunächst eine Varianzanalyse durch Anpassung des ANOVA Modells vorgenommen. Für das resultierende Modell stellte die *glht*-Funktion allgemeine lineare Hypothesentests auf. Hierbei erfolgte der Vergleich der Durchschnittswerte paarweise für alle Datensätze unter Akzeptanz der unterschiedlichen Größen der Datensätze. Des Weiteren nahm die Funktion keine Normalverteilung der Daten an.

Kapitel 3

Ergebnisse

3.1 Evaluierung des cgMLST Schemas in EnteroBase

Bevor die cgMLST Analyse zur Lösung epidemiologischer Fragestellungen angewendet wurde, die Typisierbarkeit und das Auflösungsvermögen des cgMLST Schemas evaluiert. Hierfür wurden alle verfügbaren 13.515 Genome in EnteroBase in die Analyse einbezogen.
Das cgMLST Schema in EnteroBase zeigte eine Typisierbarkeit von 99,96 % und erfasste für die 13.515 untersuchten Genome durchschnittlich 2.555 der 2.556 cgMLST Loci. Dabei traten bei zwei Einträgen mit 1159 und 1096 nicht vorhandenen Allelen auffällig hohe Lücken in den zugehörigen Genomen auf. Eines dieser Genome gehörte der Spezies *Clostridioides mangenotii* an und eines fiel in die weiter unten näher beschriebenen kryptischen Kladen.

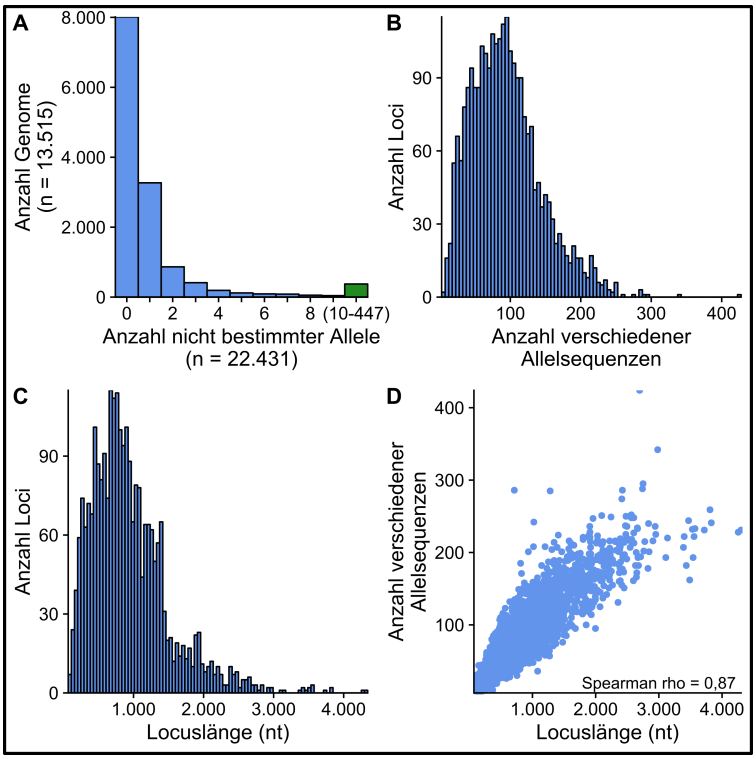

Abbildung 3.1: Evaluierung des cgMLST Schemas in EnteroBase. **(A)** Verteilung der Anzahl an unbestimmten Allelen pro Genom. Der grüne Balken fasst die Ausreißer zusammen. **(B)** Verteilung der Anzahl an unterschiedlichen Allelsequenzen pro Locus. **(C)** Verteilung der Sequenzlängen pro Locus. **(D)** Einfluss der Sequenzlänge auf die Alleldiversität pro Locus.

Im Durchschnitt konnten für jedes Genom 2.554 der 2.556 cgMLST Loci einer Allelnummer zugeordnet werden (Abbildung 3.1 A). Demnach waren 0,06 % der Allelsequenzen nicht bestimmbar, was auf

Fragmentierung, Duplikation oder schlechte Qualität der Sequenzen zurückzuführen ist. Die meisten Genome in EnteroBase besaßen ein vollständiges cgMLST Allelprofil. Es gab nur vereinzelt Einträge, für deren Genome bis zu 447 Allele nicht bestimmt werden konnten. Eine durchschnittliche Anzahl von 94±50 Allelsequenzen pro Locus deutete auf die genomische Diversität der in EnteroBase erfassten *C. difficile* Isolate hin (Abbildung 3.1 B). Diese Alleldiversität stieg, wie zu erwarten, mit zunehmender Locuslänge (Abbildung 3.1 D; Tabelle 3.1, Spearman'scher Rangkorrelationskoeffizient: 0,87; $p<0,05$).

Mit der Implementierung der hierarchischen Clusterung bietet EnteroBase eine Möglichkeit zur Evaluierung des Auflösungsvermögens des cgMLST Schemas. Einträge, deren cgMLST Allelprofile sich in 0 Allelen unterscheiden, fallen in dasselbe HC0 Cluster. Die 13.515 Einträge in EnteroBase, die auch epidemiologisch zusammenhängende Genome beinhalteten, wurden in 9.186 HC0 Cluster unterteilt. Das größte HC0 Cluster bestand aus 93 Genomen. Dies schien allerdings ein Einzelfall zu sein, da ein HC0 Cluster durchschnittlich 1±2 Einträge umfasste (Simpsons Diskriminierungsindex von 0,99).

Tabelle 3.1: Statistiken der Evaluierung des cgMLST Schemas. Die Zahlen beziehen sich auf die cgMLST Profile der 13.515 Genome, die zu dem Zeitpunkt der Analyse in EnteroBase zur Verfügung standen. Die Alleldiversität beschreibt die Anzahl der verschiedenen Allelsequenzen pro Locus. Das hierarchische Cluster HC0 fasste Einträge zusammen, dessen cgMLST Profile sich in keinem der bestimmten Allele unterschieden.

	Durchschnitt	Standardabweichung	Median	Minimum	Maximum
unbestimmte Allele	2	9	0	0	447
nicht vorhandene Allele	1	16	0	0	1159
Alleldiversität	94	50	88	6	424
Locuslänge	937	563	830	74	4301
Einträge in HC0 Cluster	1	2	1	1	93

Die in EnteroBase vorhandenen Genome konnten demnach anhand des cgMLST Schemas erfolgreich mit einem hohem Auflösungsvermögen typisiert werden. Das cgMLST Schema kann somit für die Beantwortung epidemiologischer und populationsstruktureller Fragestellungen bezogen auf *C. difficile* angewendet werden.

3.2 Die Populationsstruktur von *Clostridioides difficile*

Aufgrund der umfassenden Sammlung an Genomen und der Implementierung der hierarchischen Clusterung bietet EnteroBase die Möglichkeit die Populationsstruktur von *C. difficile* zu untersuchen. Im Folgenden werden die Verknüpfungen der hierarchischen Cluster mit den verschiedenen Populationsebenen demonstriert.

3.2.1 Detektion von pandemischen Stämmen

Drei bisher publizierte Studien nutzten die Ganzgenomsequenzierung und phylogenetische Analysen zum Nachweis von Pandemien von *C. difficile*. Die meisten der Genome, die zu einer dieser berichteten Pandemien gehören, ließen sich in ein HC10 Cluster zusammenfassen (Abbildung 3.2). Als Beispiel hierfür sei die einheitliche genomische Clusterung der aus Nutztierhaltung isolierten RT078 Isolate und der pandemischen RT017 Isolate in HC10 Cluster zu nennen. Für Letztere wurde allerdings eine Aufteilung in zwei Untergruppen berichtet, welche nicht durch die HC10 Cluster widergespiegelt werden konnte [103]. Im Gegensatz dazu konnten die für die fluorchinolonresistenten RT027 Isolate berichteten Untergruppen deutlich aufgezeigt werden (FQR1 und FQR2 [15]; Abbildung 3.2).

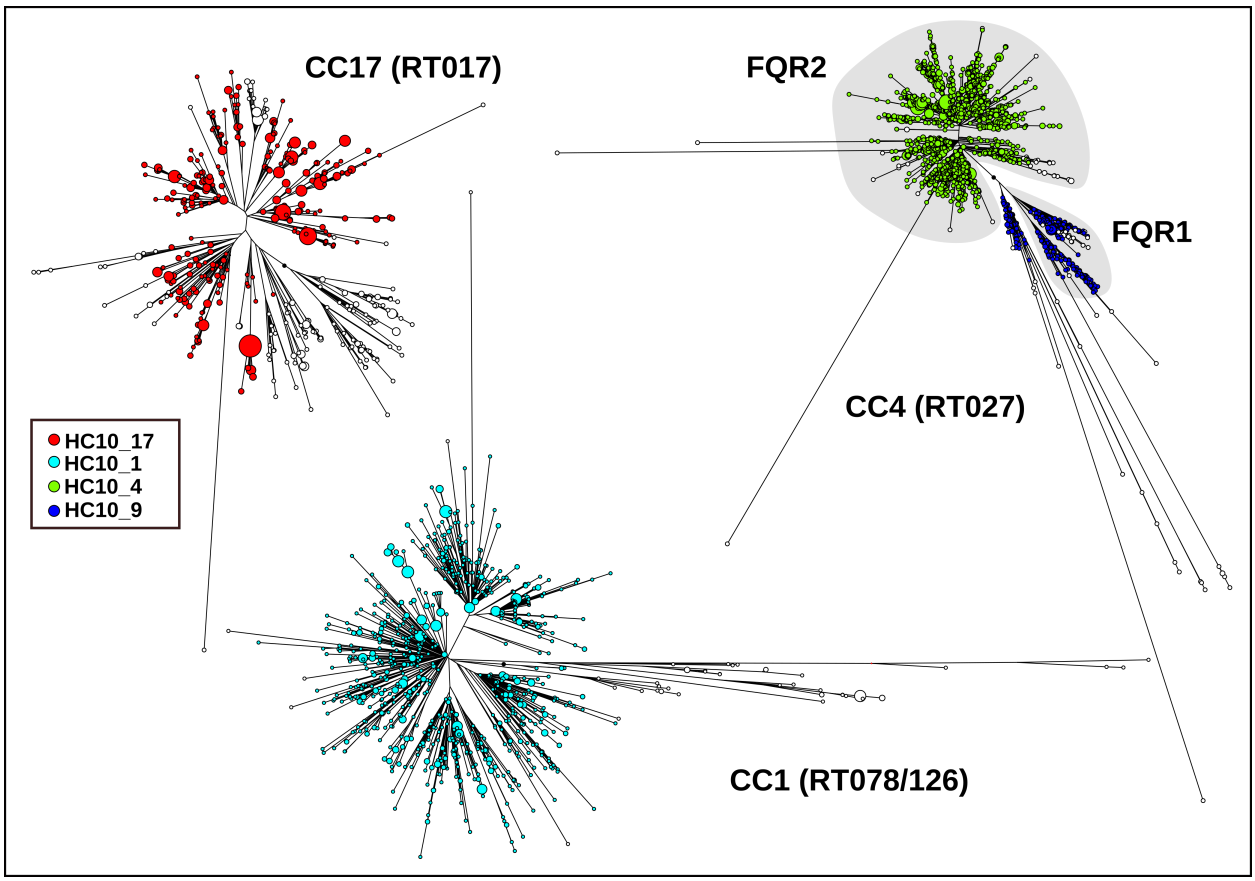

Abbildung 3.2: Phylogenetische Bäume von drei *C. difficile* Pandemien. Die Bäume wurden auf Grundlage der cgMLST Allelprofile und unter Anwendung des in EnteroBase verfügbaren Rapid-Neighbour-Joining Algorithmus berechnet. Die Farben spiegeln das jeweilige HC10 Cluster wider. CC: cgST Complex; RT: PCR Ribotyp.

Pandemische Stämme gehen meist aus endemischen Populationen hervor. Normalerweise werden solche Populationen mittels PCR Ribotypisierung charakterisiert. Eine weitere genombasierte Typisierungsmethode wird in dem folgenden Kapitel eingeführt.

3.2.2 Die hierarchische Clusterung in HC150 Cluster - eine Alternative zur PCR Ribotypisierung

Die PCR Ribotypisierung verwendet Primer für die konservierten 16S rRNA und 23S rRNA Regionen, um die dazwischenliegende intergenische Spacer-Region zu amplifizieren. Durch die Auftrennung der dadurch entstehenden DNA-Fragmente ergeben sich individuelle Bandenmuster für endemische Stämme von *C. difficile*, welche als PCR Ribotypen bezeichnet werden. Einer der größten Nachteile der PCR Ribotypisierung ist die fehlende Standardisierung der Nomenklatur und die teilweise nicht eindeutig voneinander unterscheidbaren Bandenmuster. Mit der Implementierung einer einheitlichen Clustermethode könnte EnteroBase die Möglichkeit für eine neuartige Typisierungsmethode für endemische *C. difficile* Stämme bieten. Durch das per Definition örtlich begrenzte Auftreten von endemischen Stämmen wurde eine genomische Verwandtschaft zwischen zugehörigen Isolaten angenommen. Eine Häufigkeitsverteilung der Kerngenom-Allelunterschiede aller paarweisen Genomvergleiche in EnteroBase zeigte, dass ein Großteil der Genome sich in weniger als 150 Kerngenom-Allelen unterschied (Abbildung 3.3). Da sich diese Genome in HC150 Cluster zusammenfassen ließen, wurde das HC150 Cluster als potentielle alternative Typisierungsmethode ausgewählt.

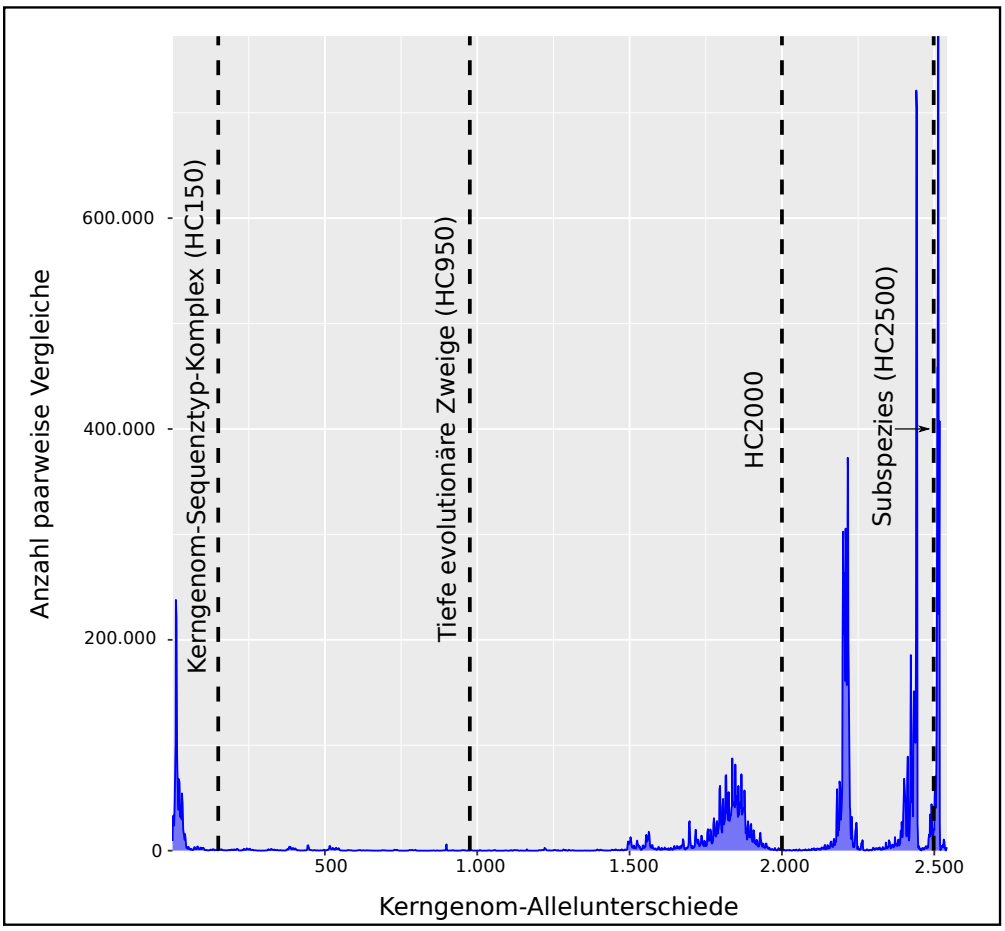

Abbildung 3.3: Häufigkeitsverteilung der Kerngenom-Allelunterschiede zwischen den verfügbaren 13.515 Genomen in EnteroBase. Die Linien geben die Grenzen der hierarchischen Cluster HC150, HC950, HC2000 und HC2500 an den entsprechenden Kerngenom-Allelunterschieden an.

Eine Literaturrecherche ergab, dass, zusammen mit den im Rahmen dieser Arbeit ribotypisierten Stämmen, für 2.263 Einträge in EnteroBase die Information des PCR Ribotypen zur Verfügung stand. Eine phylogenetische Analyse der zugehörigen Genome resultierte größtenteils in einer Übereinstimmung der phylogenetischen Gruppen mit den PCR Ribotypen. Allerdings wurden auch Defizite der PCR Ribotypisierung aufgezeigt. Zum einen wurden Genome, dessen cgMLST Allelprofile sich nicht oder nur geringfügig unterschieden verschiedenen Ribotypen zugeordnet (Abbildung 3.4 A; z.B. RT001/RT241, RT027/RT176, RT106/500). Zum anderen zeigten Genome, die dem gleichen Ribotypen zugeordnet wurden, deutliche Unterschiede in ihren cgMLST Allelprofilen (Abbildung 3.4 A; z.B. RT002, RT018). Hier sei neben den Genomen, die sich trotz gleichem PCR Ribotypen in phylogenetisch weit voneinander entfernten Clustern befanden, besonders die Genome des PCR Ribotyps 014 zu nennen. Zwar fielen diese in phylogenetische Gruppen, die sich auf einem Ast befanden, wurden aber durch die hierarchische Clusterung in vier verschiedene HC150 Cluster gruppiert. Es gab also endemische Stämme, deren Identifikation durch PCR Ribotypisierung nicht eindeutig möglich war. Des Weiteren führte die nicht einheitliche Nomenklatur zur gleichen Bezeichnung von genomisch diversen Stämmen. Betrachtet man allerdings die HC150 Cluster, stimmte die Clusterbildung eindeutig mit den phylogenetischen Gruppen überein (Abbildung 3.4 B). Trotz der aufgezeigten Diskrepanzen zeigten die beiden Methoden grundsätzlich eine hohe Übereinstimmung (Adjusted Rand Koeffizient, 0,92; 95 % Konfidenzintervall, 0,90-0,93).

Aufgrund der aufgezeigten Defizite der PCR Ribotypisierung schlägt diese Arbeit die Verwendung der HC150 Cluster als neue Typisierungsmethode für *C. difficile* Genome vor. Im weiteren Verlauf dieser Arbeit werden die zur Typisierung verwendeten HC150 Cluster mit 'CC' abgekürzt (Kerngenom-Sequenztyp-Komplex; cgST Complex).

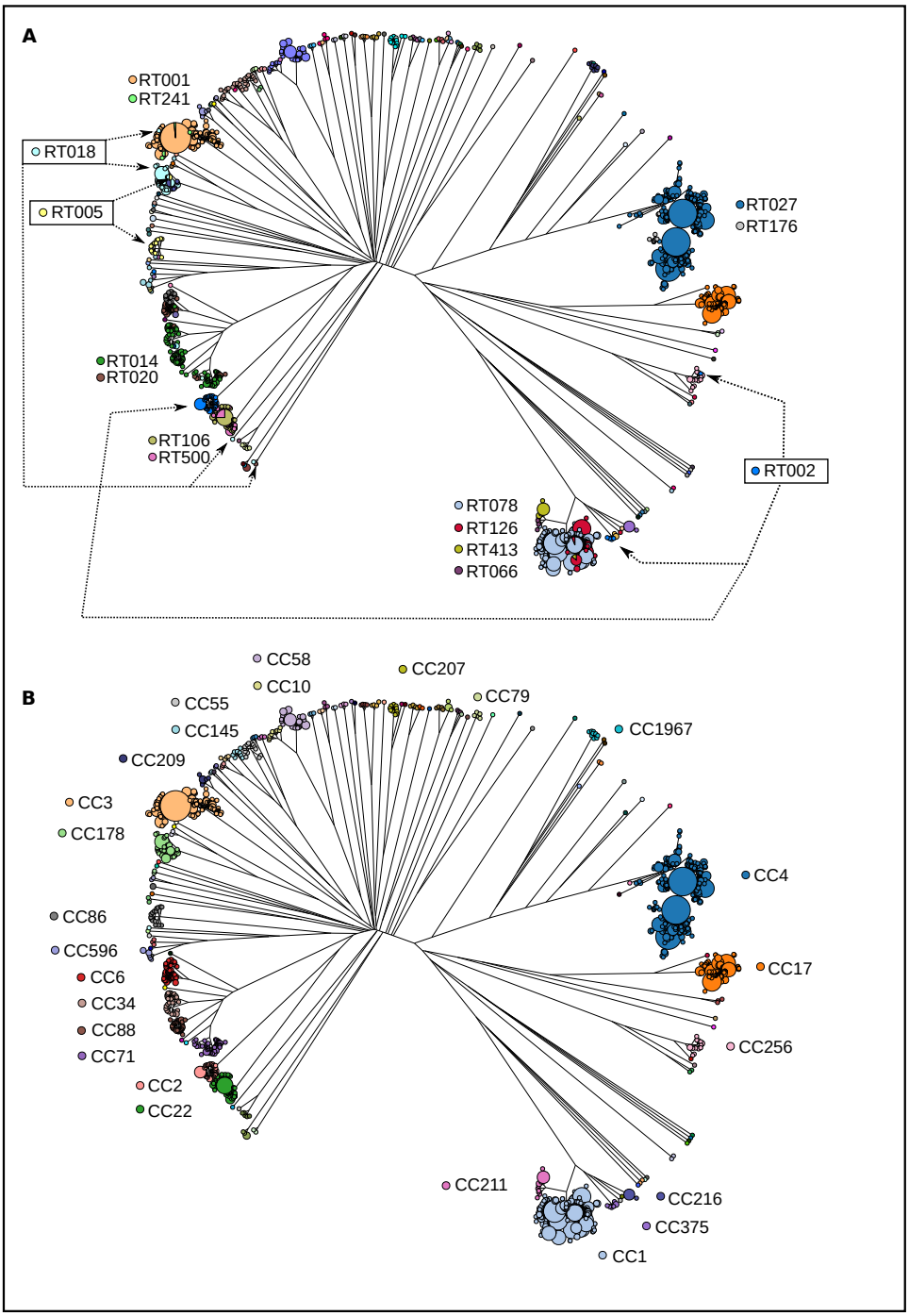

Abbildung 3.4: Vergleich der HC150 Cluster und PCR Ribotypen. Rapid-Neighbor-Joining phylogenetische Bäume basierend auf cgMLST Allelprofilen von 2.263 Genomen, für die die Information über den PCR Ribotypen zur Verfügung stand. **(A)** Farben zeigen PCR Ribotypen an. **(B)** Farben zeigen CCs an. CC: cgST Complex.

Der in Abbildung 3.5 dargestellte phylogenetische Baum zeigt die Einteilung aller 13.515 Einträge in insgesamt 201 CCs. Diese CCs setzten sich deutlich von den so genannten kryptischen Kladen und der nah verwandten Spezies *C. mangenotii* ab, die hier als Wurzel des Baumes fungierte. Neben einer Vielzahl von kleineren, phylogenetisch distinkten Gruppierungen, ließen sich die bekannten Ribotypen wie RT027 (CC4), RT078/106 (CC1) und RT001 (CC3) eindeutig erkennen.

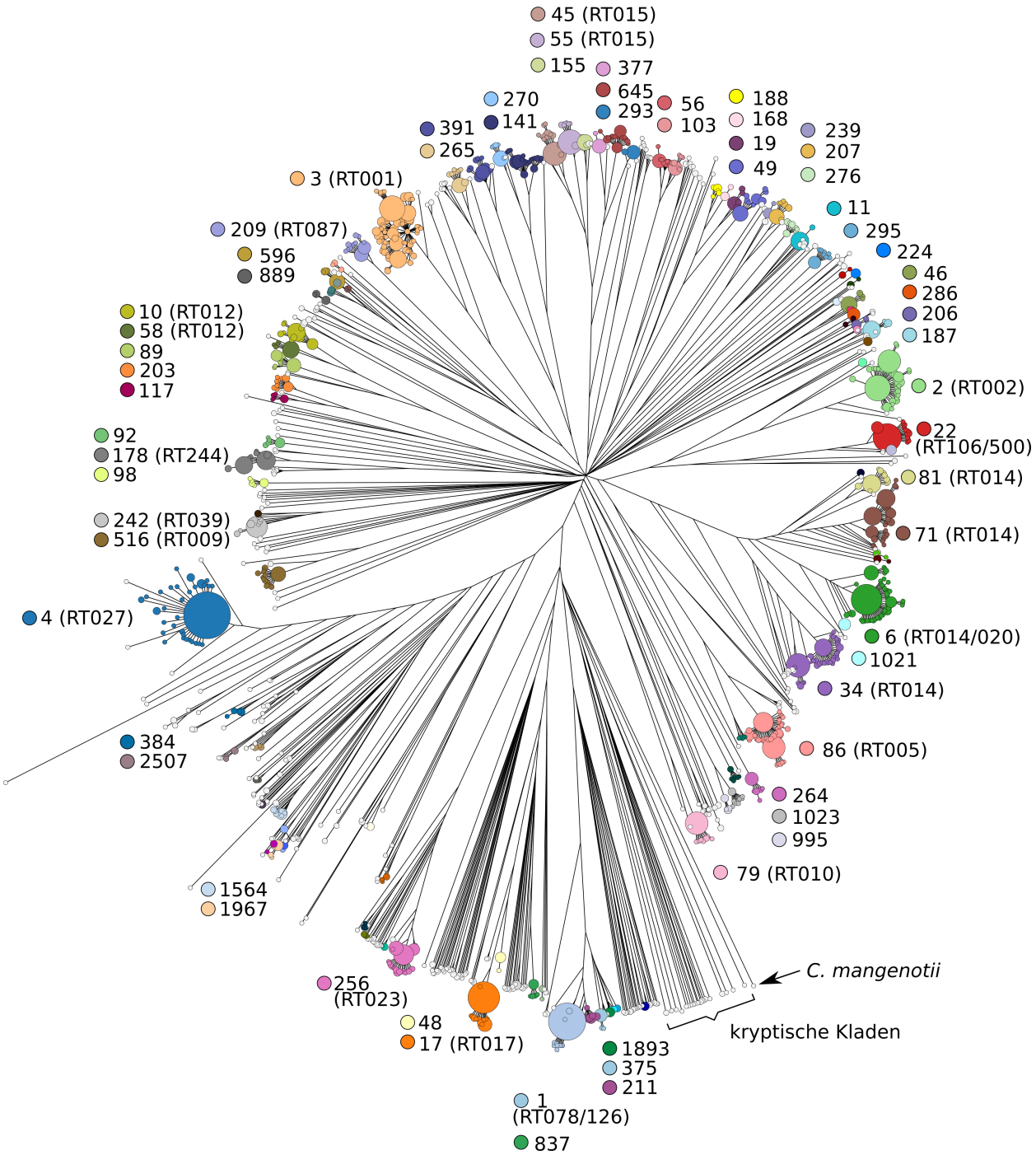

Abbildung 3.5: Phylogenetischer Baum aller 13.515 *C. difficile* Genome in EnteroBase, basierend auf deren cgMLST Allelprofilen. Zur Berechnung des Baumes wurde der Rapid-Neighbour-Joining Algorithmus verwendet. Die Farben und Zahlen stehen für die entsprechenden CCs, die mindestens zehn Einträge umfassen. Zugehörige PCR Ribotypen werden in den Klammern erwähnt. CC: cgST Complex; RT: PCR Ribotyp.

Zusätzlich bildeten sich 209 Singletons, welche sich genomisch in >150 Kerngenom-Allelunterschieden von allen anderen Isolaten in EnteroBase unterscheiden. Ein großer Teil der Singletons befand sich auf dem gleichen Ast wie CC4 und bildete mit diesem ein Monophylum. Die Bildung eines Monophylums mit Singletons ließ sich auch für die CCs 256, 17, 48 und 837 beobachten.

Die in EnteroBase zur Verfügung stehenden Metadaten zeigten, dass die meisten der größeren CCs über viele Jahre identifiziert wurden. Besonders hervorzuheben sind hier CC3, CC4 und CC86, welche bis zu 37 Jahren bestehen blieben (Tabelle 3.2). Es wurde zudem deutlich, dass alle CCs in mehr als einem Land vorkamen.

Tabelle 3.2: Übersicht über CCs mit ≥ 100 Einträgen. Die Information über Länder, Isolationsjahr und Quelle der Isolate wurde aus den in EnteroBase verfügbaren Metadaten gewonnen.

CC (HC150)	PCR Ribotyp	Anzahl an Einträgen	Anzahl an Ländern	Isolationsjahr	% der Isolate aus Tieren
4	027	2.669	27	1985–2018	0
1	078,126	1.222	26	1994–2018	17
17	017	769	24	1990–2017	0
3	001	768	16	1980–2017	0
6	020,404	768	14	1995–2017	1
2	002	702	15	2006–2017	1
22	106,500	531	7	1997–2017	3
86	005	468	8	1980–2017	0
34	014	421	10	1995–2017	0
55	015	318	6	2006–2017	0
71	014,020	315	16	2004–2017	1
145	015	284	7	2006–2016	0
256	023	268	6	2001–2015	0
79	010	249	7	2003–2018	3
178	018,356	243	7	2006–2017	0
242	039	199	4	2008–2017	1
10	012	159	7	1996–2017	0
88	014	132	9	1996–2016	8
11	070	110	6	2006–2017	0
187	054	109	6	2007–2018	0
141	001,026	107	2	2007–2016	0
391	018	105	4	1996–2016	0
49	011,056,446	103	5	2001–2017	0

Cluster, die pandemische Stämme aus Kapitel 3.2.1 enthielten, breiteten sich sogar in bis zu 27 Ländern aus. Außerdem stach der hohe Anteil an tierischen Isolaten in CC1 heraus. Zusammen führten die aufgezeigten Punkte zu der Annahme, dass *C. difficile* Stämme über eine lange Zeit ein genetisch stabiles Kerngenom aufzeigten und über große geographische Distanzen verbreitet wurden. Die Anzahl von 201 CCs, in die sich die verfügbaren Genomsequenzen einteilen ließen, spiegelte nur zwei Drittel der tatsächlich existierenden Diversität der *C. difficile* Genome wider (Abbildung 3.6). Sogar bei der Einteilung in HC950 Cluster, welche eine kettenweise genomische Distanz von bis zu 950 Allelunterschieden zulässt, schien es immer noch nicht-repräsentierte Cluster zu geben. Ein phylogenetischer Baum auf cgMLST Ebene aller verfügbaren Genome in EnteroBase deutete an, dass das HC950 Cluster CCs mit demselben Vorfahren zusammenfasst (Anhang B Abbildung B.1) und mit tiefen evolutionären Verzweigungen übereinstimmt. Erst bei den hierarchischen Cluster HC2000 und HC2500 stieg die Extrapolation nicht mehr an, hier wurden scheinbar alle zu unterscheidenden Gruppen erfasst. Die HC2000 Cluster zeigten eine Übereinstimmung mit den großen fünf *C. difficile* Kladen, die im Jahre 2014 publiziert und seitdem nicht erweitert wurden [152] (Anhang B Abbildung B.2).

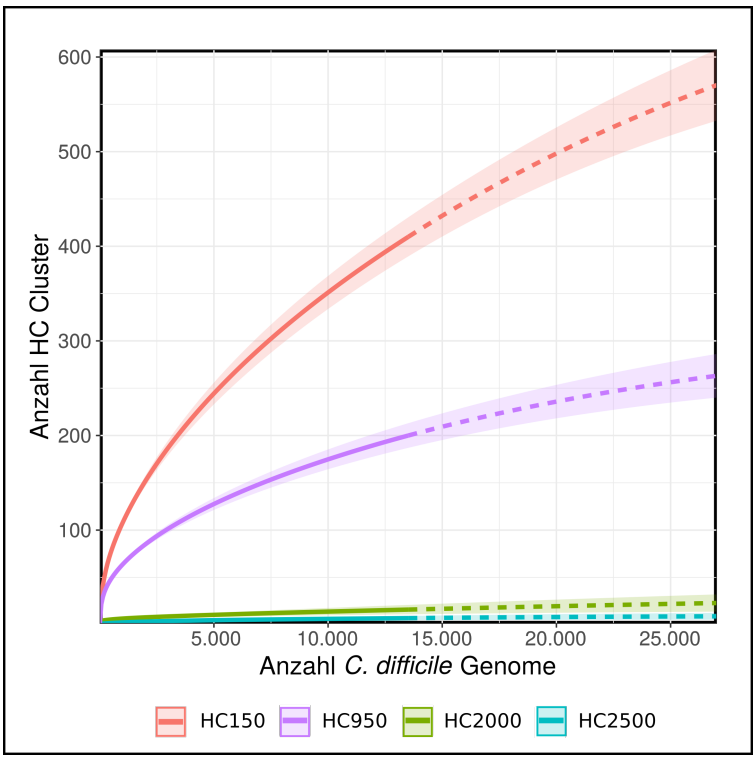

Abbildung 3.6: Rarefaction Analyse der hierarchischen Cluster HC150, HC950, HC2000 und HC2500. Die gestrichelten Linien stellen die Extrapolation und damit die Schätzung der vorhandenen hierarchischen Cluster über die schon in EnteroBase erfassten dar.

Allerdings wurden die Kladen 1 und 2 in ein HC2000 Cluster zusammengefasst und nicht wie beschrieben als individuelle Cluster detektiert. Als fünfte Klade bildete sich eher eine Gruppe, die hauptsächlich aus Singletons besteht, zwischen den bekannten Kladen 3 und 4 aus (Anhang B Abbildung B.2; HC2000_21). Bei einem kettenweisen genomischen Unterschied von bis zu 2500 cgMLST Allelunterschieden wurden bis auf die kryptischen Kladen und die Spezies *C. mangenotii* alle verfügbaren Genome in EnteroBase in ein HC2500 Cluster zusammengefasst (Anhang B Abbildung B.3). Dies zeigte sich auch in der hohen Anzahl an paarweisen Distanzen von >2500 cgMLST Allelunterschieden (Abbildung 3.3).

3.2.3 Beispielanwendung der Typisierung durch EnteroBase

EnteroBase wurde zur Typisierung von drei Isolaten aus städtischen Gewässern verwendet. Die Isolate waren Teil einer Studie von Numberger et al., die den ersten Nachweis von *C. difficile* in Badeseen und Kläranlagen in Berlin erbringt [90]. Durch die Sequenzierung der Genome und das Hochladen der Illumina-*Reads* auf EnteroBase konnten die Isolate einem CC zugeordnet werden (Tabelle 3.3). Dadurch konnten nahe genomische Verwandtschaften zu anderen Genomen in der EnteroBase Datenbank detektiert werden. Hierbei wurde eine nahe genomische Beziehung zu einem Genom von einem menschlichen Isolat aus England festgestellt. Dies ist erneut ein Hinweis auf die Verbreitung eines *C. difficile* Stammes über Landesgrenzen hinaus und dessen Vorkommen in unterschiedlichen Wirten.

Tabelle 3.3: Analyse von drei *C. difficile* Wasserisolaten. Die Information über die nächsten Verwandten wurden durch die verfügbaren Metadaten in EnteroBase bestimmt. CC: cgST Complex; NA: nicht verfügbar.

Isolat	nächstes verwandtes Isolat in EnteroBase	CC	verwandter PCR Ribotyp	Herkunftsland	Isolationsquelle
DSM104450	CLO_AA6355AA	CC071	RT010	NA	NA
DSM104451	CLO_AA5670AA	CC295	NA	Vereinigtes Königreich	Human
DSM104452	CLO_AA4563AA	CC079	RT014	Deutschland	Human

Der erste Teil der Ergebnisse demonstrierte wie die in EnteroBase implementierte hierarchische Clusterung bei der Übersichtsgewinnung der Populationsstruktur von *C. difficile* helfen kann. Es wurde gezeigt, wie pandemische Stämme mit Hilfe des HC10 Clusters detektiert werden können und das sich bestimmte *C. difficile* Typen über lange Zeit in mehreren Ländern verbreiten. Es ist jetzt interessant herauszufinden, ob die hierarchische Clusterung zur Detektion von direkten Transmissionswegen genutzt werden kann. Folglich wird sich das nächste Kapitel mit dem Vergleich zwischen der auf dem cgMLST Schema in EnteroBase basierten cgMLST-Analyse und der standardmäßig zur Detektion von Transmissionswegen verwendeten Analyse beschäftigen: Der SNP-Analyse.

3.3 Quantitativer Vergleich der cgMLST- und SNP-Analyse

Um Ausbrüche oder direkte Transmissionswege von *C. difficile* aufzudecken, wird oft die SNP-Analyse herangezogen. Dabei wird meist auf eine Publikation von Eyre et al. aus dem Jahre 2013 verwiesen die zeigt, dass Isolate, die einen kettenweise genomischen Unterschied von ≤2 SNPs aufzeigen, mit einer 95 % Wahrscheinlichkeit einer Transmissionskette zugeordnet werden können [72]. Allerdings sind SNP-Analysen rechenintensiv und fordern ein hohes Maß an bioinformatischem Verständnis. Zudem stehen eine Vielzahl an bioinformatischen Werkzeugen zur Verfügung, die zu unterschiedlichen Ergebnissen führen können.

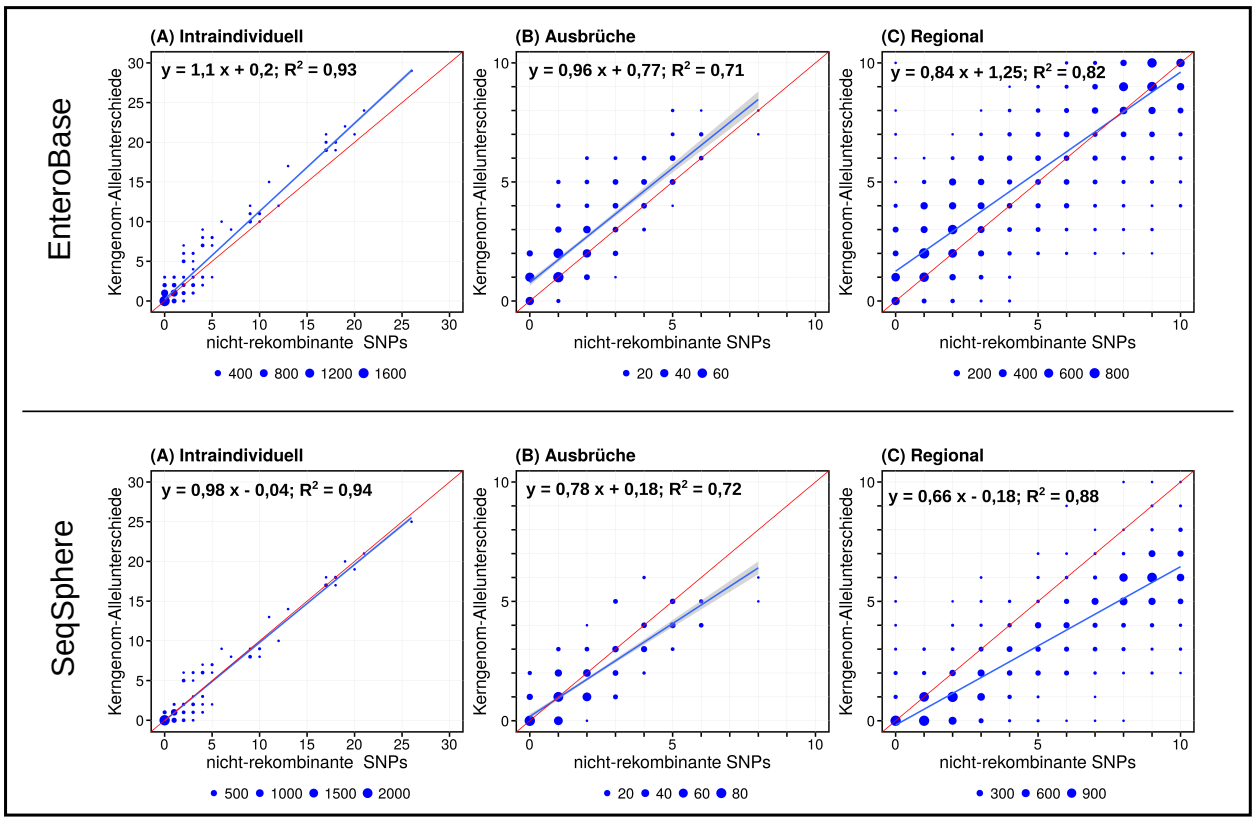

Abbildung 3.7: Korrelationsanalyse zwischen den genomischen Distanzen der cgMLST- und SNP-Analyse. Die oberen Graphen basieren auf den cgMLST Allelprofilen aus EnteroBase, die unteren auf denen aus SeqSphere$^+$. Die Größe der Punkte spiegelt die Anzahl der Datenpunkte wider. Die Graphen zeigen die Gegenüberstellung der paarweisen Kerngenom-Allelunterschiede und SNP Distanzen für die folgenden Datensätze: **(A)** Isolate von vier rezidivierenden CDI Patienten, welche an zwei Zeitpunkten beprobt wurden (Anzahl der Isolate, n = 176). **(B)** Isolate von vier kürzlich publizierten Ausbrüchen, darunter ein Ausbruch in einem chinesischem Krankenhaus, der sich über zwei Jahre zog (n = 12) [112] und ein Ausbruch in Süddeutschland, der zwei Krankenhäuser betraf (n = 9) [120]. Zudem beinhaltet der Datensatz noch zwei Ausbrüche aus einem Krankenhaus in Madrid (Spanien), welche die PCR Ribotypen 027 (n = 22) und 106/500 (n = 20) betreffen [55]. **(C)** Isolate, die zwischen 2007 und 2011 in vier Krankenhäusern in Oxfordshire, Vereinigtes Königreich, von CDI erkrankten Patienten isoliert wurden (n = 1.158) [72]. Die genomischen Distanzen werden hier bis zu einem Wert von 10 gezeigt.

Die cgMLST-Analyse könnte durch die intuitive Anwendung und einheitliche Prozessierung der Sequenzdaten eine Alternative zur Identifizierung von Transmissionsketten bieten. Auch die kommerziell verfügbare Software SeqsSphere$^+$ beinhaltet ein eigens entwickeltes cgMLST Schema für *C. difficile*. Dieses soll im Folgenden zusammen mit dem in EnteroBase verfügbaren Schema mit der SNP-Analyse verglichen werden. Die Anwendung eines einfachen linearen Regressionsmodells auf die auf cgMLST- und SNP-Analysen basierenden genomischen Distanzen resultierte in einer starken linearen Abhängigkeit (R^2, 0,71-0,93; Abbildung 3.7). Die Steigung ließ auf einen 1:1 Anstieg der Kerngenom-Alllelunterschiede mit den SNP Distanzen schließen. Die gleiche Analyse wurde für die paarweisen Distanzen basierend auf dem cgMLST Schema in SeqsSphere$^+$ durchgeführt. Obwohl die Korrelation ähnliche Abhängigkeiten andeutete (R^2, 0,71-0,94), schien die Steigung der Kurven deutlich niedriger zu sein.

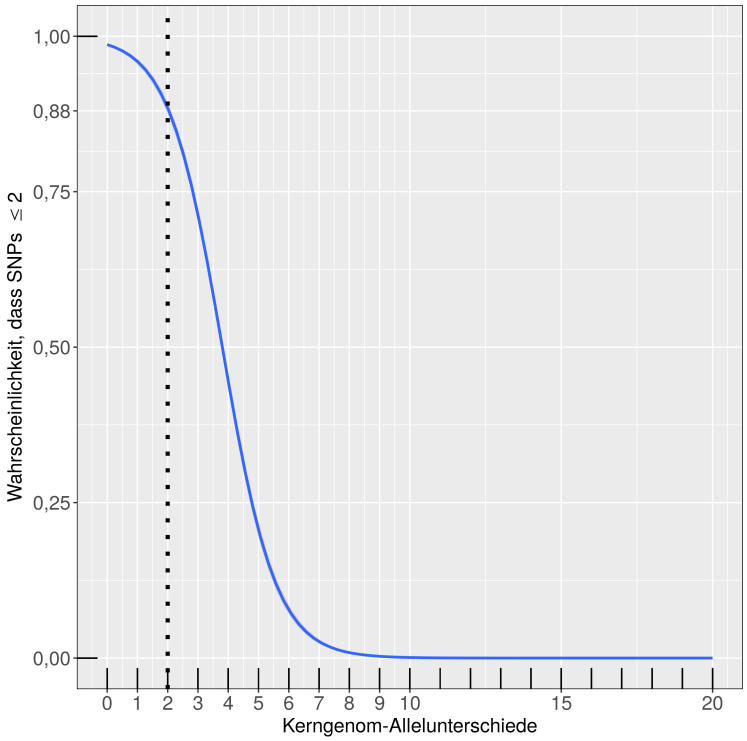

Abbildung 3.8: Binäres logistisches Regressionsmodell angewandt auf den Oxfordshire Datensatz von 1.158 Genomen. Um die Wahrscheinlichkeit zu berechnen, dass zwei Genome sich bei einem gewissen Kerngenom-Allelunterschied in ≤2 SNPs unterscheiden, wurden die SNP Distanzen binär codiert (1 wenn ≤2 SNPs, 0 wenn >2 SNPs). Die Allelunterschiede wurden als Vorhersagevariabel genutzt.

Um herauszufinden, ob die cgMLST-Analyse vergleichbare Ergebnisse wie die SNP-Analyse bei der Detektion von Transmissionsketten erzielt, wurde ein binäres logistisches Regressionsmodell auf den Datensatz aus Oxfordshire, Bestandteil der zuvor erwähnten Publikation von Eyre et al.[72], angewendet. Die Untersuchung ergab, dass mit einer 89 %igen Wahrscheinlichkeit Genompaare mit einer Distanz von ≤2 SNPs auch durch ≤2 Kerngenom-Allelunterschiede detektiert worden wären (95 % Konfidenzintervall: 88-89 %). Eine 100 %ige Wahrscheinlichkeit wurde nicht erreicht, was auf Diskrepanzen in den quantitativen Vergleichen der Distanzen der beiden Methoden hindeutet. Dies äußert sich in der <0,1 %igen Chance, dass die cgMLST-Analyse eine genomische Distanz von ≤2 SNPs vorhersagt, die beiden Genome sich aber in Wirklichkeit in ≥2 SNPs unterscheiden. Endgültig ausschließen ließ sich eine genomische Distanz von ≤2 SNPs bei einem Wert von >9 Kerngenom-Allelunterschieden.
Abschließend führte eine Untersuchung des Datensatzes von Eyre et al. auf Transmissionsketten durch die cgMLST-Analyse zu Ergebnissen, die mit den publizierten, auf der SNP Analyse beruhenden Werten vergleichbar sind [72]. Von den 1.158 analysierten Genomen wurden 916 während der tatsächlichen Testperiode isoliert (April 2008–März 2011). Die verbleibenden 242 Isolate wurden in einer Zeit von sechs Monaten vor der Testperiode isoliert, sodass Pathogene, die eine Infektion ausgelöst haben könnten,

auch erfasst wurden. Insgesamt zeigten 35 % (318/916) der Isolate eine genomische Distanz von ≤2 Kerngenom-Allelunterschieden zu einem früheren isolierten Isolat. In der Studie von Eyre et al. wurde der Zusammenhang für 34 % (316/916) der Isolate durch die SNP-Analyse erbracht. Beide Methoden identifizierten zu 89 % die gleichen Isolate.
Die dargestellten Ergebnisse in diesem Kapitel zeigen, dass die cgMLST-Analyse vergleichbare Ergebnisse wie die SNP-Analyse erzielt und somit zur Untersuchung epidemiologischer Fragestellungen angewendet werden kann. Folglich können sequenzbasierte Detektionen von Transmissionswegen auf eine standardisierte Art und ohne große bioinformatische Vorkenntnisse durchgeführt werden.

3.4 Anwendung der cgMLST-Analyse auf lokale und globale Epidemiologie

EnteroBase bietet mit der hierarchischen Clusterung eine Möglichkeit zur Einteilung der Genome auf verschiedenen Ebenen. Zur Detektion von Transmissionswegen sind genomische Distanzen von ≤2 Kerngenom-Allelunterschieden von Bedeutung, da hiermit vergleichbare Ergebnisse erzielt wurden wie mit der standardmäßig verwendeten SNP-Analyse (siehe Kapitel 3.3). Isolate, die sich kettenweise genomisch in ≤2 Kerngenom-Allele unterscheiden werden in einem HC2 Cluster zusammengefasst. Aufgrund dessen werden in den folgenden Kapiteln HC2 Cluster zur Detektion von Transmissionsketten in mehreren Beispieldatensätzen herangezogen.

3.4.1 Differenzierung eines Rezidivs von einer Neuinfektion durch *C. difficile*

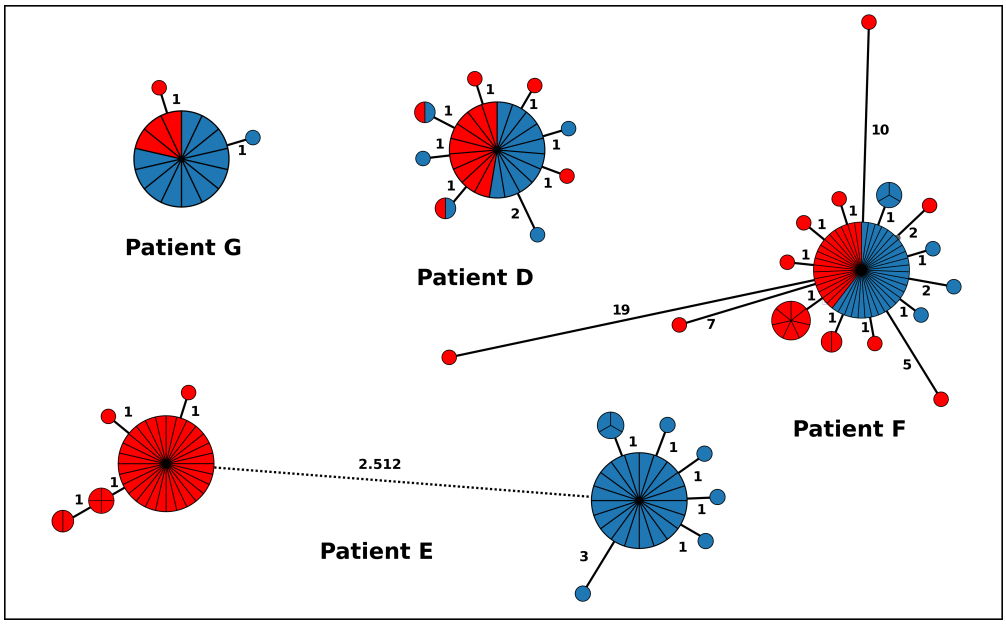

Abbildung 3.9: Wiederkehrende C. difficile Infektionen. Minimum-Spanning Bäume basierend auf den cgMLST Allelprofilen der Genome von Isolaten, die aus vier Patienten mit wiederkehrender CDI zu zwei Zeitpunkten isoliert wurden. Die Zahlen geben die Anzahl der cgMLST Allelunterschiede an. Rot: Erster Aufenthalt; blau: Zweiter Aufenthalt.

Durch eine cgMLST-Analyse von *C. difficile* Isolaten, isoliert aus vier Patienten, welche zwei Krankenhausaufenthalte aufgrund einer *C. difficile* Infektion (CDI) hatten, konnte zwischen einer Neuinfektion und einer rezidivierenden Infektion unterschieden werden. Eine wiederkehrende Infektion, die durch den gleichen Stamm verursacht wurde, wurde durch das Clustern aller Genome in ein HC2 Cluster detektiert, ungeachtet des Zeitpunktes der Infektion (Abbildung 3.9, Patient D und G). Die nahe genomische

Verwandtschaft der Isolate deutete darauf hin, dass der Stamm die erste Behandlung überstand und den Patienten weiterhin kolonisierte.
Eine CDI kann allerdings auch durch mehrere, nah verwandte Stämme verursacht werden, was in Abbildung 3.9 von Patient F zu sehen ist. Hier wurden die Genome in mehrere HC2 Cluster aufgeteilt und zeigten Distanzen zwischen 12–21 paarweisen cgMLST Allelunterschieden.
Als deutlicher Kontrast ist hier der Fall von Patient E zu sehen. Hier teilten sich die Genome episodenweise in ein HC2 Cluster auf. Die HC2 Cluster unterschieden sich in >2000 cgMLST Allelen, was die Aufnahme eines anderen, genomisch distinkten Stammes andeutete, der die zweite Infektion verursachte. Interessanterweise lagen die Episoden der rezidivierenden Patienten 16 bis 22 Wochen auseinander, was deutlich über den empfohlenen Grenzwert zur Detektion von rezidivierenden Erkrankungen von acht Wochen liegt [169].

3.4.2 Anwendung der hierarchischen Clusterung zur Detektion von lokalen und regionalen Ausbrüchen von *C. difficile*

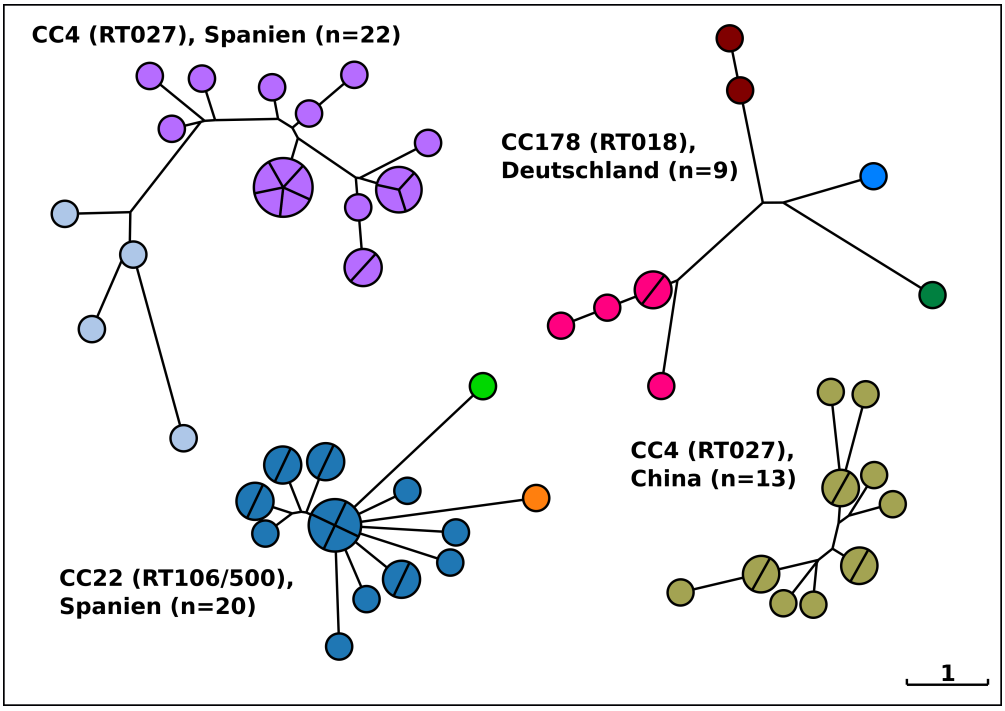

Abbildung 3.10: Phylogenetische Bäume von vier publizierten Ausbrüchen, basierend auf Kerngenom-Allelunterschieden [55], [112], [120]. Die Bäume wurden mit dem Rapid-Neighbour-Joining Algorithmus berechnet. Die Farben indizieren die HC2 Cluster. Die Skala, die für einen Kerngenom-Allelunterschied steht, gilt für alle Bäume. CC: cgST Complex; RT: PCR Ribotyp.

Aufgrund der bisherigen Ergebnisse wäre zu erwarten, dass die Genome von Isolaten, die im Zuge eines Ausbruches isoliert wurden, in ein HC2 Cluster fallen. Die phylogenetischen Bäume von vier publizierten Ausbrüchen zeigten allerdings, dass dies nicht der Fall war (Abbildung 3.10). Beispielhaft hierfür war die Clusterung der Genome des RT018 Ausbruchs in Süddeutschland in vier HC2 Cluster. Dieser Ausbruch breitete sich über zwei Krankenhäuser aus, was zu Lücken in der Erfassung von infizierten Patienten geführt haben könnte und somit zu fehlenden Verknüpfungen in der Transmissionskette. Die Genome der Isolate der in Madrid erfassten Ausbrüche der Ribotypen 027 und 106/500 wurden in zwei beziehungsweise drei HC2 Cluster sortiert. Diese Studie erfasste alle diagnostizierten CDI Fälle während der Zeit der Studie. Hier könnten die fehlenden Verknüpfungen auf asymptomatische Patienten zurückzuführen sein, welche routinemäßig nicht auf Kolonisierung durch *C. difficile* untersucht wurden.

Genome von Isolaten, die im Zuge eines Ausbruchs isoliert wurden, fallen also nicht zwangsweise auch in ein HC2 Cluster. Das eine einheitliche Clusterung möglich ist zeigt das Beispiel des China-Ausbruchs. Hier wurden alle Isolate in einem HC2 Cluster erfasst.

Eine Analyse der genomischen Einteilung der in einem Netzwerk von Krankenhäusern isolierten Isolate in HC2 Cluster resultierte in einer retrospektiven Detektion von multiplen Transmissionsketten. 46 % der Genome der Isolate aus der Studie wurden in insgesamt 23 HC2 Cluster gruppiert (133 von 309 Isolate). Dabei umfasste das größte Cluster 66 Genome. Durch die Analyse der zugehörigen epidemiologischen Daten ließ sich eine Assoziation zwischen der Bildung von HC2 Clustern und einzelnen Krankenhäusern nachweisen (χ^2, p=8,6×10^{-5}; Shannon Entropie, p=4,2×10^{-5}). Diese Assoziation konnte auch für einzelne Stationen gezeigt werden (χ^2, p=0,01; Shannon Entropie, p=6,2×10^{-3}), was die Hypothese der Repräsentation von lokalen Ausbrüchen durch HC2 Cluster unterstützt.

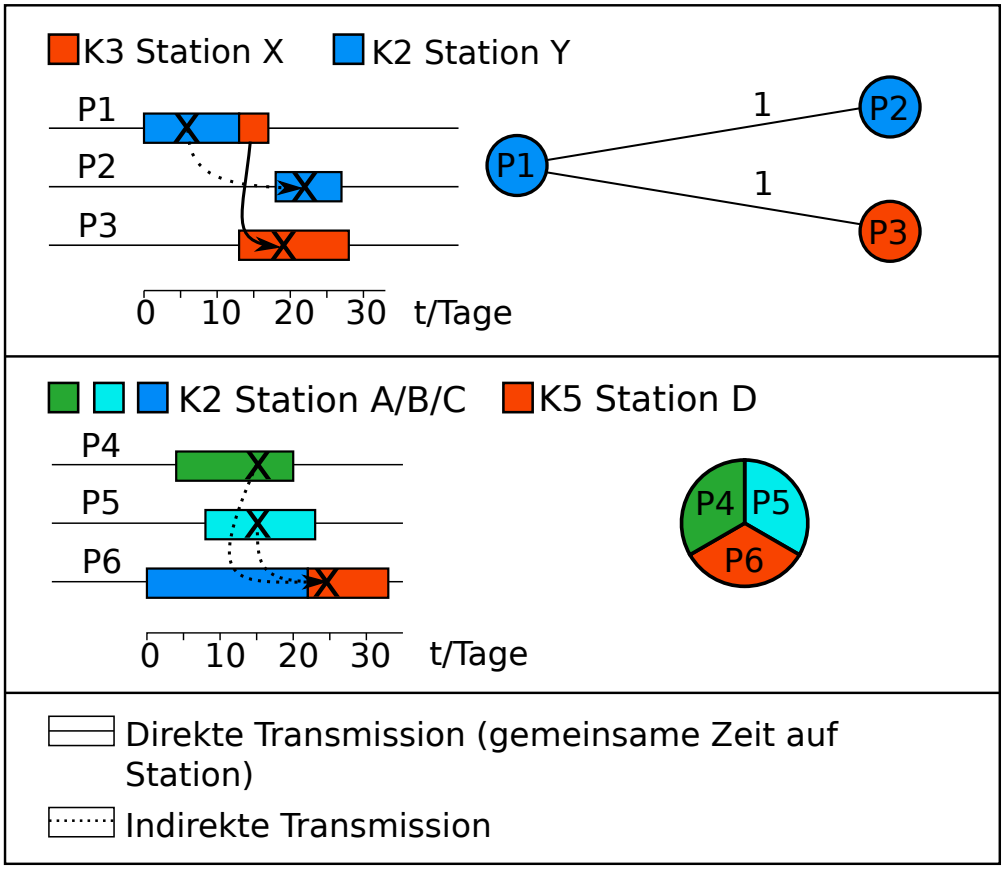

Abbildung 3.11: Nachgewiesene Transmissionswege in einem Netzwerk von Krankenhäusern. Die Farben geben die Stationen wider, 'X' steht für den Zeitpunkt der Diagnose der CDI und die Pfeile indizieren die angenommenen Transmissionswege. Rechts befinden sich die zugehörigen Minimum-Spanning Bäume, welche die genomischen Distanzen zwischen den Isolaten zeigen. Oberes Feld: Patient P1 wurde in Krankenhaus K2 mit CDI diagnostiziert und fünfzehn Tage später in Krankenhaus K3 verlegt. Fünf Tage nach der Verlegung wurde in Krankenhaus K3 Patient P3 mit einem nah verwandten Stamm infiziert, ebenso Patient P2 in Krankenhaus K2 nach sechs Tagen. Beide Patienten lagen im jeweiligen Krankenhaus auf der gleichen Station wie der Erstpatient, welcher wahrscheinlich die Infektionsquelle darstellt. Allerdings gab es bei den Patienten P1 und P2 keine zeitliche Überschneidung, sodass die Übertragung wahrscheinlich über die Umgebung stattfand. Unteres Feld: Die Patienten P4 und P5 wurden am selben Tag mit CDI diagnostiziert, nachdem sie sieben Tage lang gemeinsam hospitalisiert waren, allerdings auf verschiedenen Stationen. Bei einem dritten Patient P6 wurde eine CDI mit einem genomisch identischem Stamm wenige Zeit später diagnostiziert. Obwohl die Diagnose in einem anderen Krankenhaus stattfand, liegt eine mögliche Transmission vor, da Patient P6 vorher im gleichen Krankenhaus wie die Patienten P4 und P5 hospitalisiert war.

Insgesamt wurde für 66 Patienten, deren Isolate eine genomische Verwandtschaft von ≤2 Kerngenom-Alleluntерschieden zu einem Isolat eines anderen Patienten aufwiesen, ein Aufenthalt auf der gleichen Station nachgewiesen. Hier wurde zwischen direkten und indirekten Transmissionen unterschieden. Bei direkten Transmissionen lagen die Patienten zur gleichen Zeit auf einer Station (Abbildung 3.11 A; Patient P1 und P3), bei indirekten gab es keine zeitliche Überschneidung der Aufenthalte der Patienten (Abbildung 3.11 A; Patient P1 und P2). Hierbei lagen die Aufenthalte der beiden Patienten bis zu 521 Tage auseinander (Durchschnitt: 63 Tage). Indirekte Transmissionen konnten auch zwischen Patienten nachgewiesen werden, die zwar nicht auf der gleichen Station, aber im gleichen Krankenhaus lagen (Abbildung 3.11 B; Patient P4/P5 und P6). Am überraschendsten war allerdings der Nachweis von 15 HC2 Clustern, die Genome von Isolaten aus verschiedenen Krankenhäusern beinhalteten. Für manche konnte nach Sichtung der epidemiologischen Daten ein Zusammenhang mit Patientenverlegungen hergestellt werden (Abbildung 3.11 B). Besonders für diese Fälle wäre die Erkennung eines Zusammenhangs durch die standardmäßig durchgeführte Surveillance schwer geworden. Dies traf auch auf die Detektion von drei rezidivierenden CDI zu. In einem dieser Fälle wurden die beiden Episoden in unterschiedlichen Krankenhäusern diagnostiziert.
Neben den nahen genomischen Verwandtschaften zwischen Isolaten innerhalb dieser Studie, wurde diese auch zu anderen Genomen in EnteroBase nachgewiesen. In 15 der 23 HC2 Cluster befanden sich neben den hier analysierten Genomen noch insgesamt 79 Genome ohne epidemiologischen Bezug. Detailliertere Ergebnisse hierzu werden in dem Kapitel 3.5 gezeigt.

Bis jetzt wurden nosokomiale Transmissionswege von *C. difficile* Isolaten aufgezeigt. Dass *C. difficile* Isolate sich aber möglicherweise auch durch die Ausbringung von tierischem Dung auf landwirtschaftlich genutzte Flächen ausbreiten können, wird mit den folgenden Ergebnissen demonstriert.

3.4.3 Verbreitung von *C. difficile* durch Ausbringung von tierischem Dung auf landwirtschaftlich genutzte Flächen

Das Projekt SOARiAL untersuchte die mögliche Übertragung von antibiotikaresistenten und krankheitserregenden Keimen durch die Ausbringung von Dünger aus der Tierhaltung auf landwirtschaftlich genutzte Flächen in den Boden und durch die Bearbeitung sowie Erosion über die Luft. Im Zuge dessen wurden auch *C. difficile* Isolate isoliert, sowohl aus dem Dünger und aus Sammelkotproben, als auch aus dem Boden des Testfeldes vor und nach Ausbringung des Düngers. Die Genome der Isolate wurden im Zuge dieser Arbeit mit der cgMLST-Analyse ausgewertet.
Betrachtet man die Verteilung der CCs in die sich die Genome clustern ließen, fällt auf, dass der im Mist dominante CC3 sich bis zu 19 Wochen in dem gedüngten Boden und in dem Staub wiederfinden ließ, dieser allerdings nicht in dem ungedüngten Boden auftrat (Abbildung 3.12). Die vor der Ausbringung des Dungs isolierten *C. difficile* Isolate aus dem Boden ließen sich einheitlich dem CC103 zuordnen.
Insgesamt umfasste der Datensatz 191 Genome, die sich in 15 HC2 Cluster einteilen ließen. Neun Genome zeigten keine nahe Verwandtschaft zu einem anderen Genom in EnteroBase und bildeten Singletons. Da sich die vor der Düngung isolierten Bodenisolate schon auf CC-Ebene nicht mehr in den anderen Proben wiederfanden, separierten sie sich auch hier in eigene HC2 Cluster Abbildung 3.12). Überraschend war allerdings, dass zwei der 15 HC2 Cluster Genome aus EnteroBase beinhalteten, welche keinen direkten epidemiologischen Zusammenhang zu den SOARiAL Isolaten hatten (Abbildung 3.13).
Ein Isolat aus dem Boden, das zwei Wochen nach Düngerausbringung isoliert wurde, zeigte eine nahe genomische Verwandtschaft zu fünf Krankenhausisolaten (Kapitel 3.4.2; Abbildung 3.13, HC2 Cluster 1206). Dieser Zusammenhang wurde auch in dem HC2 Cluster 1232 gefunden. Hier umfasste das Cluster zwei Drittel der Genome von Isolaten aus dem SOARiAL Projekt (118 Genome), wobei Isolate aus Mist, gedüngtem Boden und Staub vertreten waren. Das Cluster umfasste damit alle Probenquellen mit Ausnahme von ungedüngtem Boden. Ein Großteil der Genome war auf Kerngenom-Ebene identisch, was darauf schließen

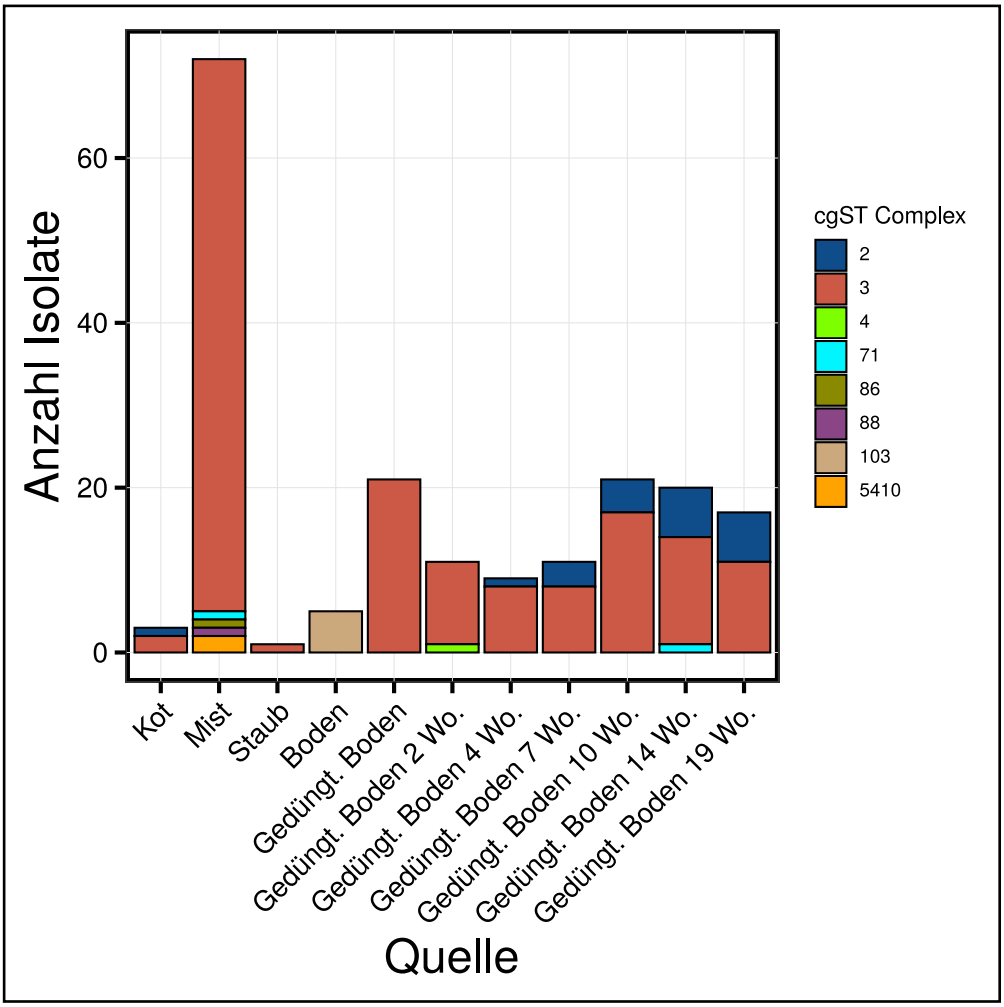

Abbildung 3.12: Verteilung der Isolate in den Proben des SOARiAL Projekts (n = 191). Die Farben geben die den Genomen in EnteroBase zugeordneten CCs an. Die Wochenangaben der gedüngten Bodenproben geben die Zeit nach Ausbringung des Düngers auf den Boden an.

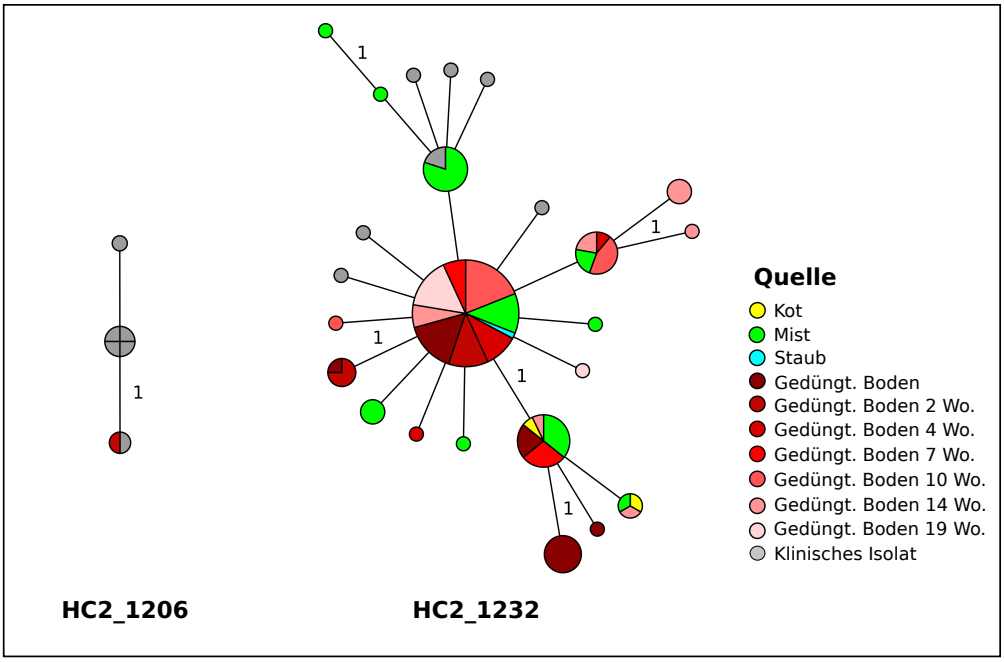

Abbildung 3.13: Minimum-Spanning Bäume der HC2 Cluster 1206 und 1232, die sowohl Genome von Isolaten aus dem SOARiAL Projekt als auch Genome von klinischen Isolaten in EnteroBase umfassten. Die Farben indizieren die Isolationsquelle des zugehörigen Isolats.

ließ, dass Isolate in dem ausgebrachten Dung bis zu 19 Wochen auf dem Feld überlebten. Neben der nahen genomischen Verwandtschaft zu einem Krankenhausisolat aus Kapitel 3.4.2 ließ sich dies auch zu einem humanen Isolat aus Ungarn nachweisen. Des Weiteren umfasste das HC2 Cluster fünf humane Isolate aus den Niederlanden, welche laut der paarweisen Vergleiche der Genome die engste Verwandtschaft zu Isolaten aus dem Dung aufzeigten. Diese Zusammenhänge deuteten auf eine mögliche ubiquitäre Verbreitung einiger *C. difficile* Isolate hin.
Auch wenn nur ein Isolat aus dem Staub, der bei der Durchführung des Windkanalversuchs aufgewirbelt wurde, gewonnen werden konnte, so ist es doch interessant zu sehen, dass dieses ein zu den Isolaten aus Mist und gedüngtem Boden identisches Kerngenom besaß (Abbildung 3.13, HC2 Cluster 1232). Dieser Sachverhalt deutete auf eine mögliche Verteilung von *C. difficile* Isolaten durch den entstehenden Staub bei Ausbringung des Dungs hin.

Somit wurden sowohl für Isolate aus dem SOARiAL Datensatz, als auch für Isolate aus den regionalen Ausbrüchen eine genomische Distanz von $\leq$2 Kerngenom-Allelunterschieden zu Isolaten nachgewiesen, obwohl aufgrund der Herkunft und des Probezeitpunktes kein offensichtlicher epidemiologischer Zusammenhang ersichtlich war. Daher galt es anschließend die gesamte Datenbank nach solchen Fällen zu untersuchen.

3.4.4 Analyse der globalen Verbreitung von *C. difficile*

Frühere Studien deckten Transmissionen von *C. difficile* über Landesgrenzen hinaus und zwischen unterschiedlichen Wirtsspezies anhand von SNP-Analysen auf [56], [87]. Da mithilfe der HC2 Cluster regionale Transmissionsketten aufgezeigt werden konnten, wurde diese Methode auf die ganze Datenbank in EnteroBase angewendet um hier nach möglichen globalen Transmissionswegen zu suchen.

Tabelle 3.4: Übersicht der HC2 Cluster in den größten CCs. Die Bezeichnung HC2 global steht für HC2 Cluster, die mindestens aus drei Genomen von Isolaten bestehen, welche mindestens aus zwei Ländern stammen.

CC (HC150)	Anzahl an HC2 global	Anzahl an Einträgen in HC2 global (%)	Max. Anzahl an Ländern in einem HC2 global	Anzahl an Genomen im größten HC2 global (%)	Max. Zeitdifferenz in einem HC2 global (Jahre)
4	6	844 (32 %)	9	721 (85 %)	16
1	11	395 (32 %)	9	105 (27 %)	20
17	5	40 (5 %)	6	28 (70 %)	16
3	5	83 (11 %)	3	41 (49 %)	7
6	6	55 (7 %)	4	21 (38 %)	7
2	3	40 (6 %)	4	31 (78 %)	11
22	3	126 (24 %)	2	120 (95 %)	11
86	1	16 (3 %)	3	16 (100 %)	8
55	2	16 (5 %)	3	13 (81 %)	11
71	3	22 (7 %)	2	10 (45 %)	6
145	2	10 (4 %)	3	5 (50 %)	5
79	1	4 (2 %)	2	4 (100 %)	5
178	4	35 (14 %)	3	14 (40 %)	6
10	3	21 (13 %)	4	13 (62 %)	12
88	1	5 (4 %)	2	5 (100 %)	4
11	1	3 (3 %)	2	3 (100 %)	1
187	1	17 (16 %)	2	17 (100 %)	10

Für 17 der 23 größten CCs aus Tabelle 3.2 ließen sich HC2 Cluster, welche mindestens drei Genome von Isolaten aus mindestens zwei Ländern umfassten, detektieren (HC2 global). Dabei stach CC1 mit einer Anzahl von insgesamt elf globalen HC2 hervor. Das größte HC2 Cluster beinhaltete allerdings nur 27 % aller Genome in globalen HC2 Clustern (Tabelle 3.4). Die anderen CCs wiesen für die längste globale HC2-Kette deutlich höhere Prozentanteile aller Genome in globalen HC2 auf. Dies deutete auf wiederholte globale

Ausbrüche des CC1 hin, die zwar nicht viele Isolate umfassten, dafür aber über lange Zeit bestehen blieben (Maximale Zeitspanne: 20 Jahre; Tabelle 3.4). Im Gegensatz dazu deutete der hohe Anteil an Genomen aus den globalen HC2 in der längsten HC2 Kette von CC4 und CC17 auf eine große Pandemie hin. Dabei sei darauf hingewiesen, dass die zugehörigen Isolate der längsten HC2 Kette von CC4 in insgesamt zwei Ländern isoliert wurden, die aus CC17 in sechs Ländern. Somit schien sich CC17 über mehrere Länder auszubreiten, während CC4 eine große Pandemie in zwei Ländern verursachte.
Tatsächlich ließ sich eine nahe genomische Verwandtschaft von ≤2 Kerngenom-Allelunterschieden zwischen Isolaten aus unterschiedlichen Ländern für 1.004 Genome nachweisen (9 % aller Genome in EnteroBase mit Länderinformation; Abbildung 3.14; durchschnittlich 13±20 Kerngenom-Alleleunterschiede). Im Gegensatz dazu zeigten diese nahe Verwandtschaft 6.909 Genome, die von Isolaten aus dem gleichen Land stammen (62 % aller Genome in EnteroBase mit Länderinformation; durchschnittlich 5±13 Kerngenom-Alleleunterschiede). Dieser Zusammenhang war für aus einem Land stammende Isolate zu erwarten, da die Wahrscheinlichkeit, dass diese im Zuge einer Studie isoliert wurden hoch ist. Die Anzahl der Isolate, die die nahe genomische Verwandtschaft zu einem Isolat aus einem anderen Land zeigten ist überraschend, da hier kein offensichtlicher epidemiologischer Zusammenhang vorliegt.

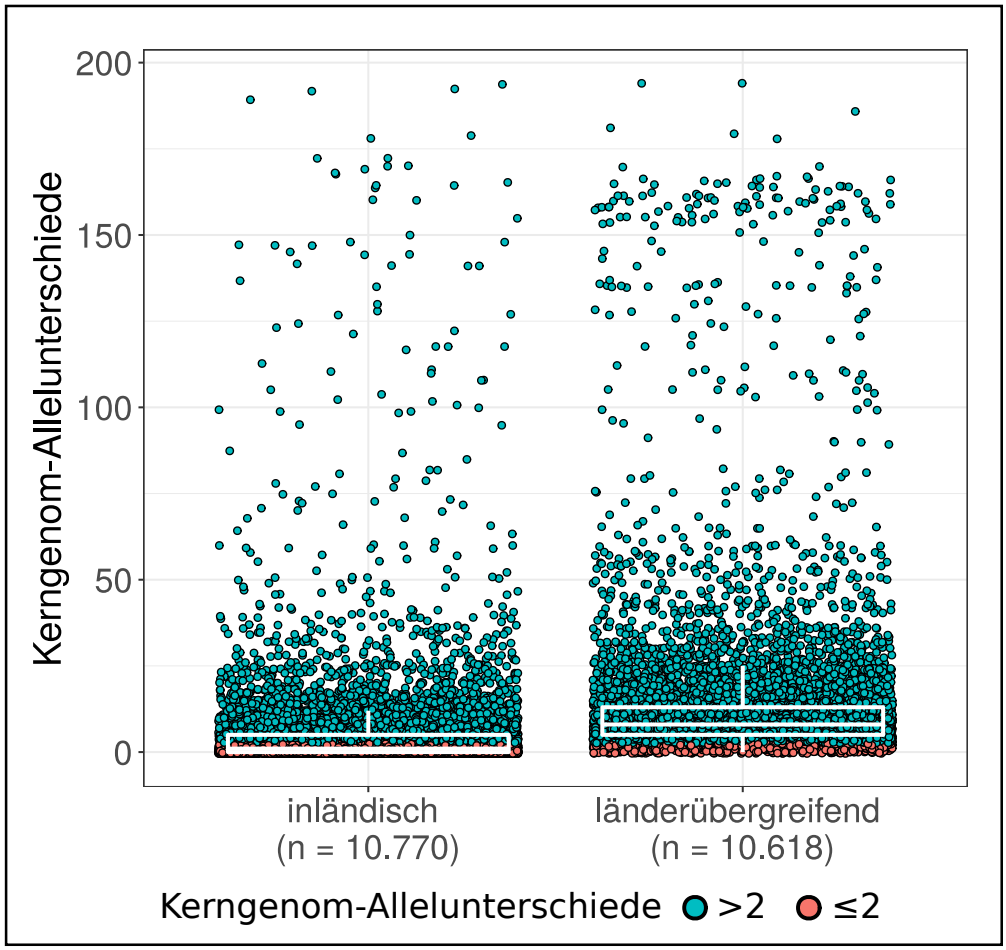

Abbildung 3.14: Kleinster Kerngenom-Allelunterschied für jedes Genom in EnteroBase mit Länderinformation zu einem anderen Genom aus dem gleichen Land („inländisch“) und zu einem Genom aus einem anderen Land („länderübergreifend“), dargestellt bis zu einer Distanz von <200 (n = 10.957). Die Farbe gibt an, ob dieser Wert ≤2 (rot) oder >2 (blau) Kerngenom-Allelunterschiede beträgt. Die Boxplots geben Information über Median, unteres und oberes Quartil.

Um diese nahe genomische Verwandtschaft zu bestätigen, wurden für die 1.004 Genome die paarweisen SNP Distanzen mithilfe der SNP-EToKi-Analyse berechnet. Überraschenderweise ließ sich ein genomischer Zusammenhang von ≤2 SNPs zu einem Isolat aus einem anderen Land nur für 50 % der 1.004 Genome bestätigen. Dies widerspricht den vorangegangenen Analysen lokaler und regionaler Daten, die in einer 89 % Übereinstimmung der SNP- und cgMLST-Analysen resultierten. Da sich die SNP Distanzen nicht nur auf das Kerngenom begrenzten, waren die Mutationen, die die Diskrepanz zwischen den beiden Methoden hervorriefen, außerhalb dieser Region zu vermuten. Mit dem implementierten wgMLST Schema in EnteroBase war es möglich das akzessorische Genom auf Unterschiede zu untersuchen. Um die Analyse des akzessorischen Genoms nicht nur auf die im wgMLST Schema repräsentierten Stellen im Genom zu beschränken, wurden die Assemblierungen aus EnteroBase zusätzlich mit der Software PopPUNK analysiert. PopPUNK erstellt *k-mer* basiert stark reduzierte Darstellungen der Assemblierungen und definiert das Kern - und akzessorische Genom individuell für jeden Datensatz.

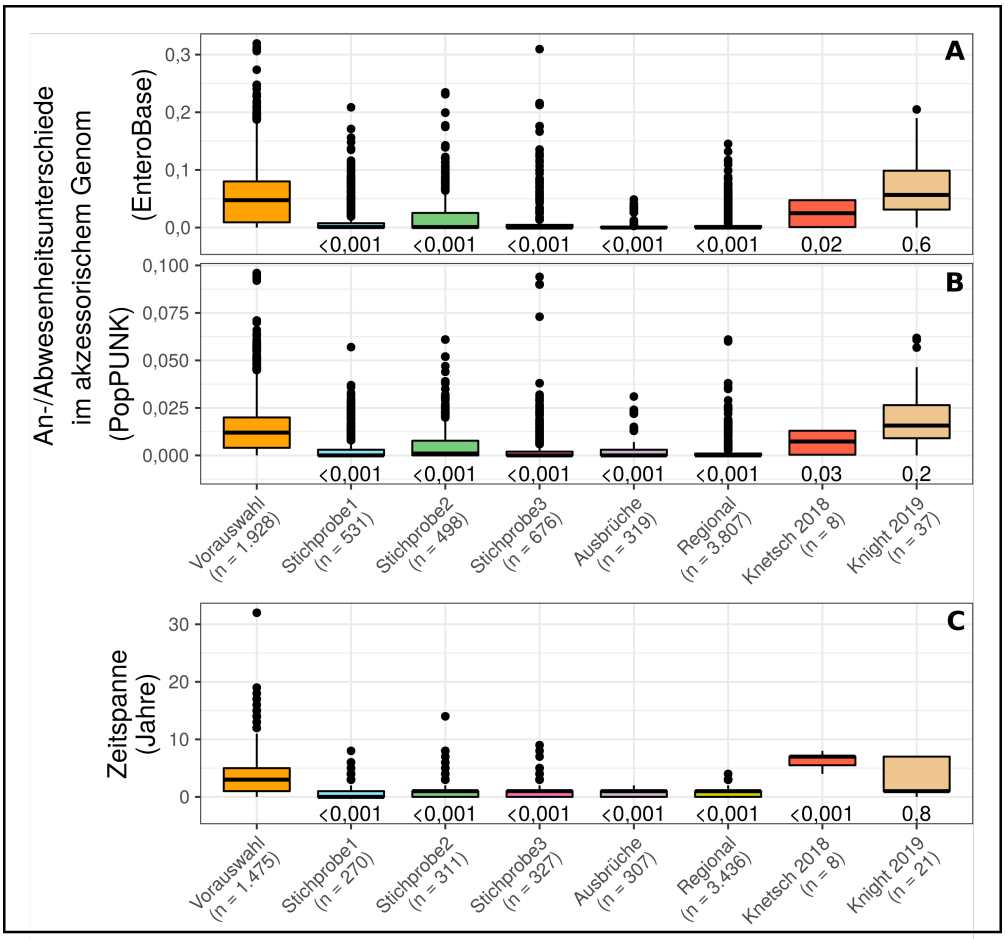

Abbildung 3.15: Boxplots der paarweisen Jaccard Distanzen im akzessorischem Genom und der Zeitspanne zwischen den Isolaten für paarweise Vergleiche mit ≤2 Kerngenom-Alllelunterschieden (Anzahl angegeben durch „n“). Vorauswahl: 1.004 Isolate, die eine nahe genomische Verwandtschaft zu einem Isolat aus einem anderen Land zeigten; Stichproben: 1.000 zufällig ausgewählte Einträge in EnteroBase für die die paarweisen Kerngenom-Alleleunterschiede bestimmt wurden; Ausbrüche: Vier publizierte Ausbrüche; Regional: Oxfordshire Datensatz aus der Publikation von Eyre et al. [72]; Knetsch 2018 und Knight 2019: Isolate, denen in der Publikation eine nahe genomische Verwandtschaft zu einem Isolat aus einem anderen Land nachgewiesen wurde [56], [87]. **(A)** Die An-/Abwesenheitsunterschiede im akzessorischem Genom wurden basierend auf den wgMLST Allelprofilen abzüglich der cgMLST Loci der Genome aus EnteroBase bestimmt. **(B)** Die An-/Abwesenheitsunterschiede im akzessorischem Genom wurden durch die *k-mer* basierte PopPUNK Analyse der Assemblierungen berechnet. **(C)** Für paarweise Vergleiche mit Jahresinformation für beide Isolate wurde die Zeitspanne zwischen den Probennahmen berechnet.

Ein Vergleich der An- und Abwesenheit der wgMLST Loci zwischen den 1.004 Genomen die auf Kerngenomebene ≤2 Alleleunterschiede zeigten, deutete auf höhere Differenzen im akzessorischem Genom hin (durchschnittliche Jaccard Distanz: 0,055; 95% Konfidenzintervall: 0,052-0,057). Dieses Verhalten ließ sich allerdings weder für genomisch nah verwandte Isolate von Ausbrüchen (0,0039; 95% Konfidenzintervall: 0,0029-0,0049) und Regionen (0,0036; 95% Konfidenzintervall: 0,0033-0,0039), noch für 1.000 zufällig ausgewählte Einträge in EnteroBase nachweisen (0,013-0,019; Abbildung 3.15 A). Die An- und Abwesenheitsdistanzen in ihren wgMLST Alleleprofilen unterschieden sich signifikant von denen in den vorausgewählten 1.004 Genomen (p-Wert <0,001). Die PopPUNK Analyse erzielte vergleichbare Ergebnisse (Abbildung 3.15 B). Weiterhin lag zwischen den Probenahmen der 1.004 vorausgewählten Isolate mit einer Alleldistanz von ≤2 im Kerngenom durchschnittlich 3,5 Jahre. Diese Zeitspanne unterschied sich signifikant von den Zeitspannen zwischen den nah verwandten Isolaten in den anderen Datensätzen (<1 Jahr, p-Wert=0,001; Abbildung 3.15 C).
Betrachtet man die Anzahl der paarweisen Vergleiche, welche im Kern- und im akzessorischem Genom den Schluss auf eine nahe genomische Verwandtschaft der Isolate zulassen, zeigte der vorausgewählte Datensatz deutlich niedrigere Werte als die anderen Datensätze (Tabelle 3.5). Die Analysen mit PopPUNK führten zu vergleichbaren Ergebnissen (Tabelle 3.5), wobei für die diversen Datensätze leicht höhere Anteile und für die epidemiologisch verwandten Datensätze leicht niedrigere Anteile detektiert wurden (∼2-5 % Abweichung; Tabelle 3.5) . Geographisch distinkte Isolate, die im Kerngenom eine nahe genomische Verwandtschaft aufwiesen, schienen sich somit im akzessorischem Genom deutlich evolutionär voneinander zu unterscheiden.

Tabelle 3.5: Anzahl an paarweisen Vergleichen ≤2 Kerngenom-Alleluntерschieden mit identischen akzessorischen Genomen bezogen auf den Gengehalt. Vorauswahl: 1.004 Isolate, die eine nahe genomische Verwandtschaft zu einem Isolat aus einem anderen Land zeigten; Stichproben: 1.000 zufällig ausgewählte Einträge aus der EnteroBase-Datenbank; Ausbrüche: Vier publizierte Ausbrüche; Regional: Oxfordshire Datensatz; Knetsch 2018 und Knight 2019: Isolate, denen in der Publikation eine nahe genomische Verwandtschaft zu einem Isolat aus einem anderen Land nachgewiesen wurde [56], [87].

Datensatz	Anzahl der paarweisen Vergleiche ≤2 Kerngenom-Allelunterschiede und 0 Ab-/Anwesenheitsunterschiede im akzessorischem Genom (%)	
	EnteroBase	PopPUNK
Vorauswahl	100 (5,19 %)	162 (8,4 %)
Stichprobe1	243 (45,76 %)	266 (50,09 %)
Stichprobe2	210 (42,17 %)	228 (45,78 %)
Stichprobe3	342 (50,59 %)	370 (54,73 %)
Ausbrüche	203 (63,64 %)	182 (57,05 %)
Regional	2.211 (58,08 %)	2.162 (56,79 %)

Eine Reanalyse der beiden erwähnten Studien am Anfang dieses Kapitels zeigte, dass diese Diskrepanz zwischen Kern- und akzessorischem Genom auch für diese Datensätze vorliegt (Abbildung 3.15 A und B: Knetsch 2018 und Knight 2019). Die Verteilung der An-/Abwesenheitsunterschiede im akzessorischem Genom unterschieden sich nicht signifikant von dem vorausgewählten Datensatz (p-Wert = 0,02; 0,9). Zumindest für den Datensatz von Knight et al. konnte eine vergleichbare Zeitspanne zwischen den genomisch nah verwandten Isolaten nachgewiesen werden (Abbildung 3.15 C; p-Wert = 0,8). Die Zeitspanne der Isolate in dem Datensatz von Knetsch et al. lag im Gegensatz zu der der anderen untersuchten Datensätze signifikant über der Zeitspanne des vorausgewählten Datensatzes (Abbildung 3.15 C; p-Wert <0,001).
Aufgrund des signifikanten Unterschiedes der Distanzen im akzessorischem Genom zu anderen, genomisch nah verwandten Isolaten, schien die nahe Verwandtschaft im Kerngenom für den vorausgewählten Datensatz nur vorgetäuscht zu sein. Da die Datensätze von Knetsch et al. und Knight et al. ähnliche Ergebnisse wie der vorausgewählte Datensatz erzielte, sind die Aussagen dieser Publikationen kritisch zu betrachten. Besonders bei Isolaten, die keinen epidemiologischen Zusammenhang haben, sollte das akzessorische Genom mit in die Analyse einbezogen werden, um keine falschen Schlüsse zu ziehen.
Dennoch konnte für 100 Isolate (0,74 % der gesamten Einträge in EnteroBase) eine nahe genomische Verwandtschaft im Kern- und im akzessorischem Genom nachgewiesen werden (Tabelle 3.5). Für diese

bestand kein offensichtlicher epidemiologischer Zusammenhang. Dieser Anteil an genomisch nah verwandten Isolaten könnte auch durch statistisches Rauschen begründet sein.
Anhand von unterschiedlichen Beispielen wurde in diesem Teil der Ergebnisse demonstriert, wie die cgMLST-Analyse und die Clusterung von Genomen in HC2 Cluster dazu beitragen haben, verschiedene epidemiologische Fragestellungen zu beantworten. Die Unterscheidung von rezidivierenden und Neuinfektionen war durch die cgMLST-Analyse möglich, genauso wie die Aufdeckung von eventuellen Transmissionswegen. Es wurde aber auch deutlich, dass der Grenzwert von ≤ 2 Kerngenom-Allelunterschieden nicht immer angewendet werden kann. Ausbrüche können sich durch asymptomatische Patienten auch über mehrere HC2 Cluster ausbreiten. Des Weiteren zeigte sich, dass eine nahe genomische Verwandtschaft im Kerngenom nicht immer eine tatsächliche Beziehung zwischen zwei Isolaten widerspiegelt. Besonders bei Genomen, die keinen offensichtlichen epidemiologischen Zusammenhang hatten, gab es deutliche Unterschiede im akzessorischem Genom. Die Anwendung von genomischen Methoden zur Untersuchung von Ausbrüchen oder genomischen Zusammenhängen ist am zuverlässigsten, wenn sie auf vordefinierte, epidemiologisch zusammenhängende Datensätze angewendet werden.

3.5 Herausforderungen der genomischen Epidemiologieanalysen

In dieser Arbeit wurden sowohl epidemiologisch zusammenhängende Genome als auch Genome ohne offensichtlichen Zusammenhang analysiert. Diese Analysen führten teilweise zu überraschenden Ergebnissen. Isolate deuteten eine genomische Verwandtschaft an, die zunächst als nicht logisch erschien und SNP- und cgMLST-Analyse widersprachen sich. Aufgrund dessen wurden die Methoden intensiv hinterfragt. Da sich ein Teil dieser Arbeit mit der Analyse von diversen Datensätzen beschäftigt und letztendlich zum Ziel hatte, die genomische Epidemiologie auch für Nicht-Bioinformatiker zugänglich zu machen, werden im Folgenden wichtige und zu berücksichtigende Punkte für die SNP- und cgMLST-Analysen aufgezeigt.

3.5.1 Nahezu identische Genome ohne epidemiologischen Zusammenhang

Eine nahe genomische Verwandtschaft von $\leq$2 SNPs wird aufgrund der von Eyre et al. durchgeführten Analysen oft als Grenzwert für einen gemeinsamen Ursprung zweier *C. difficile* Isolate in epidemiologischen Untersuchungen herangezogen[72]. Aufgrund dessen werden diese Isolate in vielen Publikationen einer Transmissionskette zugeordnet und so eine Übertragung von *C. difficile* zwischen Patienten oder unterschiedlichen Wirtsspezies vermutet [56], [87], [153]. Die Ergebnisse der vorliegenden Arbeit zeigen, dass dieser Grenzwert auch auf die cgMLST-Analyse angewendet werden konnte (Kapitel 3.3). Allerdings ist zu beachten, dass der Grenzwert nur bei epidemiologisch zusammenhängenden Datensätzen anwendbar war und diese nahe genomische Verwandtschaft nicht zum Nachweis einer möglichen Transmission ausreicht (Kapitel 3.4.4). Eine einleitende Analyse für dieses Ergebnis war eine so genannte Negativkontrolle für den in Kapitel 3.4.2 analysierten Datensatz. Die Idee dahinter war, dass wenn ein Grenzwert von $\leq$2 SNPs bzw. Kerngenom-Allelunterschieden einen epidemiologischen Zusammenhang andeutet, diese genomische Verwandtschaft zu keinem weiteren Isolat in EnteroBase außerhalb des untersuchten Datensatzes vorliegen sollte. Allerdings befanden sich in 15 HC2 Clustern, in denen mindestens eines der Genome aus Kapitel 3.4.2 fiel, ein weiteres Genom aus EnteroBase, welches nicht zu diesem Datensatz gehörte. Dies betraf insgesamt 79 Isolate in EnteroBase, wobei 23 davon ihren Ursprung außerhalb Deutschlands hatten. Die SNP-Analyse der betroffenen HC2 Cluster führte anfangs allerdings zu widersprüchlichen Ergebnissen und resultierte in deutlich höheren SNP Distanzen im Vergleich zu den Kerngenom-Allelunterschieden. Nach intensiver Analyse der einzelnen Schritte der SNP-Analyse wurden Aspekte aufgedeckt, welche die zu hohen genomischen Distanzen verursachten. Diese Aspekte werden in den folgenden Kapiteln dargestellt.

Artefakte beim Mapping Prozess

Der erste Schritt in der SNP-Analyse ist das *Read-Mapping* gegen eine Referenzsequenz. Bei der Betrachtung der *Mapping*-Ergebnisse fiel auf, dass einige Isolate an mehreren Stellen des Genoms eine ungewöhnlich hohe Abdeckung durch Illumina-*Reads* zeigten. Besonders wichtig ist hier, dass durch die hohe Abdeckung scheinbar eine Punktmutation gegenüber der Referenzsequenz nicht detektiert wurde. Zum Beispiel wurde für die Sequenzen des Isolates CD-16-00185 bei einem *Read-Mapping* mit Standardeinstellungen keine Mutation gegenüber der Referenzsequenz detektiert (Basen 49.846 und 49.851; Abbildung 3.16; A links). Dadurch unterschied es sich von den anderen Isolaten in dem hier untersuchten HC2 Cluster, die eine Mutation aufzeigten. Die bis zu zehnfach höhere Abdeckung zum Rest der alignierten *Reads* wurde durch 19 Nukleotid lange *Read*-Fragmente hervorgerufen, die an der betroffenen Stelle einen Stapel bildeten (Abbildung 3.16; B links). In diesem Stapel fand sich auch die Erklärung für die nicht detektierten Mutationen: An den Koordinaten, an denen die anderen Isolate eine Mutation gegenüber der Referenzsequenz zeigten, befand sich für das betroffene Isolat die mit der Referenzsequenz übereinstimmende Base (Abbildung 3.16; B links). Führte man den *Mapping*-Prozess mit einer *seed*-Länge von 30 durch, so wurde die gleiche Mutation gegenüber der Referenzsequenz wie in den anderen Isolaten detektiert (Basen 49.846 und 49.851; Abbildung 3.16; A rechts). Die in voller Länge alignierten Illumina-*Reads* zeigten eine Mutation gegenüber der Referenz und die 19 Nukleotid lange *Read*-Fragmente traten nicht mehr auf.

Die Auswirkungen dieser unterschlagenen Mutationen gegenüber der Referenzsequenz und dem dadurch entstehenden genomischen Unterschied zu den anderen Isolaten in dem HC2 Cluster wurde durch die Berechnung der SNP Distanzen deutlich (Tabelle 3.6). Wurde das *Read-Mapping* mit der voreingestellten *seed*-Länge durchgeführt, resultierte die SNP-Analyse noch in genomische Distanzen von bis zu >10 SNPs zwischen den auf Kerngenom-Ebene nah verwandten Isolaten. Mit der Anpassung der *seed*-Länge auf 30 Basen reduzierte sich die Anzahl der SNPs deutlich, war teilweise aber immer noch höher als die Anzahl der Kerngenom-Allelunterschiede.

Bei dem Vergleich der Koordinaten der Punktmutationen und der cgMLST Loci fiel auf, dass sich die SNPs zum Teil nicht auf den Koordinaten der cgMLST Loci befanden. Somit konnten diese SNPs nicht durch die cgMLST-Analyse detektiert werden. Es traten auch Kerngenom-Allelunterschiede zwischen Isolaten auf, die laut SNP-Analyse identisch waren (Vergleich zwischen CD-16-00166 und CLO_AA3704AA; Tabelle 3.6). Diese Unterschiede wurden durch Insertionen und Deletionen in den Allelsequenzen hervorgerufen, welche wiederum nicht in der SNP-Analyse mit einbezogen wurden. Weiterhin ist zu beachten, dass sich mehr als ein SNP auf einem cgMLST loci befinden konnte, welches in der cgMLST-Analyse nur als ein Allelunterschied detektiert wurde (Vergleich zwischen CD-16-00185 und CLO_AA3751AA; Tabelle 3.6).

Tabelle 3.6: Kerngenom und SNP Distanzen zwischen Genomen des HC2 Cluster 1.

klinisches Isolat aus Kapitel 3.4.2	Isolat aus EnteroBase	Anzahl Kerngenom-Allelunterschiede	Anzahl SNPs	Anzahl SNPs mit *seed*-Länge = 30	Anzahl SNPs auf cgMLST Loci
CD-16-00166	CLO_AA3704AA	2	12	2	0
CD-16-00166	CLO_AA3710AA	2	14	4	0
CD-16-00166	CLO_AA8571AA	2	8	3	0
CD-16-00173	CLO_AA3751AA	2	13	4	3
CD-16-00185	CLO_AA3751AA	2	12	4	3

Nachdem die SNP Distanzen mit der Einstellung der *seed*-Länge erneut berechnet wurden, konnte die nahe Verwandtschaft für viele paarweise Vergleiche in den HC2 Clustern weiterhin nicht bestätigt werden. Ein weiterer Grund hierfür schien die Wahl der Referenzsequenz zu sein, was an den folgenden Ergebnissen demonstriert wird.

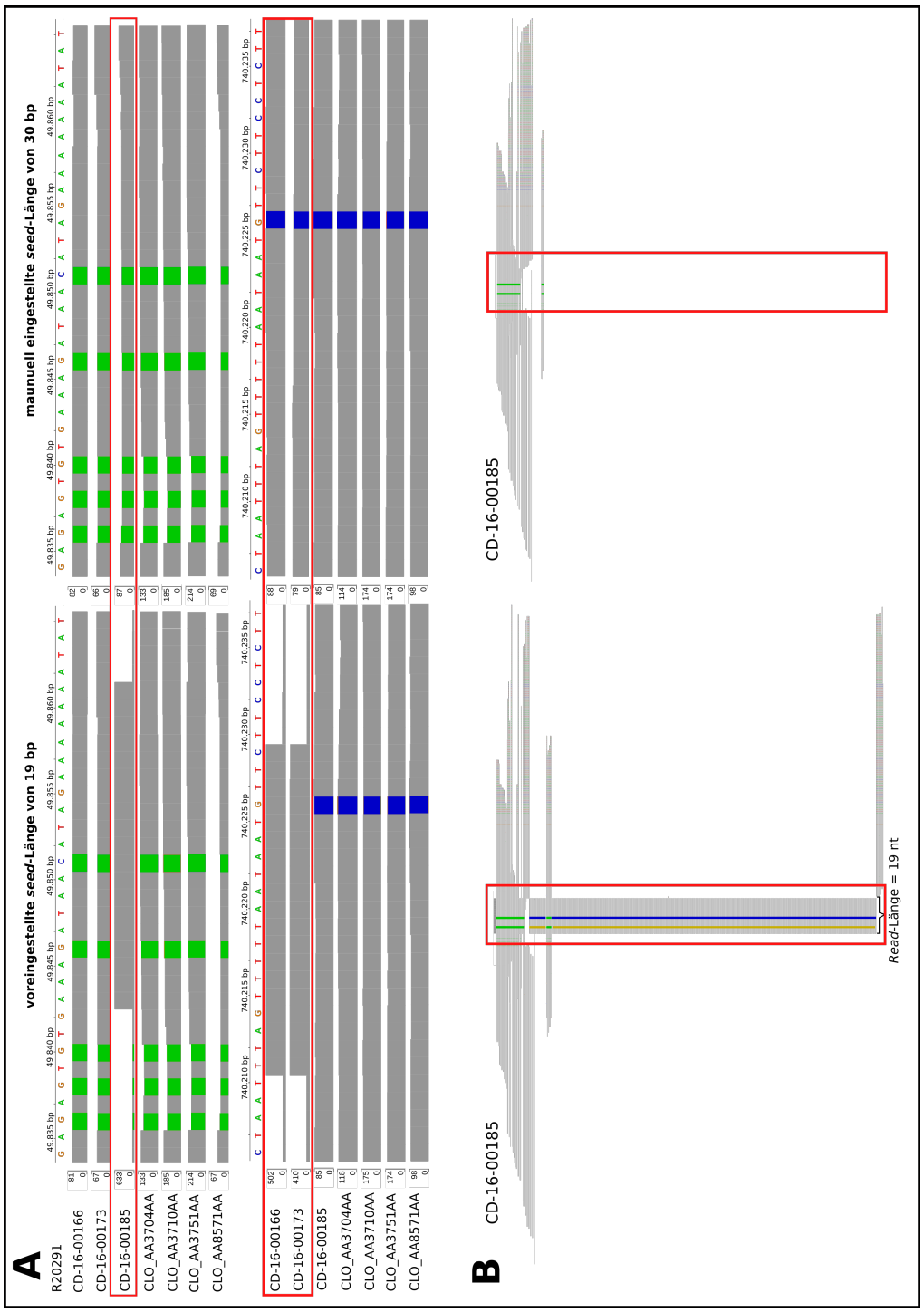

Abbildung 3.16: Darstellung von Mapping-Artefakten in den Illumina-*Reads* der Isolate von HC2 Cluster 1. Dieses Cluster umfasst sieben Einträge in EnteroBase. Die Illumina-*Reads* dieser sieben Isolate wurden gegen die Referenzsequenz R20291 gemappt. Die linken Abbildungen zeigen die Abdeckungen der Sequenzen bei einer Durchführung des *Read-Mappings* mit Standardeinstellungen (*seed*-Länge = 19), die rechten entsprechend bei einer gewählten *seed*-Länge von 30. **(A)** Nahaufnahme der Abdeckung der Illumina-*Reads* an zwei Abschnitten der Referenzsequenz, an denen SNPs zwischen den Sequenzen detektiert wurden. Die Höhe der grauen Blöcke, die das Nukleotid an dieser Stelle repräsentieren, spiegelt die Abdeckung der Illumina-*Reads* an dieser Stelle wider. Eingefärbte Blöcke zeigen einen Unterschied in der jeweiligen Sequenz zur Referenzsequenz an, wobei die Farbe der Blöcke der an dieser Stelle befindlichen DNA-Base entspricht (Adenin=grün; Guanin=gelb; Cytosin=blau; Thymin=rot). Rot markiert sind die Sequenzen, bei denen eine auffällig hohe Abdeckung detektiert wurde. Dies ist an der Skala links neben den grauen Balken zu sehen. **(B)** Darstellung der kompletten Tiefe der Illumina-*Reads* des Isolats CD-16-00185, die an den in Abbildung **A.** dargestellten Abschnitten aligniert wurden. Jeder horizontale Strich stellt einen Illumina-*Read* dar. Die rote Markierung zeigt die Bereiche des Genoms, in denen in Abbildung **A.** die auffällig hohe Abdeckung detektiert wurde. Die Farbe der vertikalen Linien entspricht der in den Illumina-*Reads* an dieser Position befindlichen Base. Die Abbildung wurde mithilfe der *IGV-Web* Anwendung erstellt (`https://igv.org/app/`).

Wahl der Referenzsequenz für den *Mapping*-Prozess

Für den *Mapping*-Prozess der Illumina-*Reads* wird empfohlen, eine Referenzsequenz zu wählen, welche eine nahe genomische Verwandtschaft zu den zu alignierenden Sequenzen hat [154]. Allerdings steht nicht immer ein vollständig sequenziertes Genom eines zu dem untersuchenden Datensatz nah verwandten Isolat zur Verfügung. Dass die Wahl der Referenzsequenz Auswirkungen auf die resultierenden SNP Distanzen hat, zeigen die Ergebnisse in Tabelle 3.7.

Tabelle 3.7: SNP Distanzen für HC2 Cluster 1 und 76, basierend auf *Read-Mappings* gegen verschiedene Referenzsequenzen. Eine Tabelle mit den Ergebnissen für alle 15 untersuchten HC2 Cluster befindet sich in Anhang B

klinisches Isolat aus Kapitel 3.4.2	Eintrag in EnteroBase	SNP Distanzen				Jahres-differenz	HC2 Cluster	wgMLST Allel-unterschiede
		R20291 (CC4)	CD196 (CC4)	CD630 (CC58)	M120 (CC1)			
CD-16-00166	CLO_AA8571AA	3	3	1	9	3	1 (CC1)	0,0648
CD-16-00166	CLO_AA3704AA	2	2	2	8	4	1 (CC1)	0,0852
CD-16-00166	CLO_AA3710AA	4	4	5	11	4	1 (CC1)	0,0724
CD-16-00173	CLO_AA3751AA	4	4	3	5	4	1 (CC1)	0,1804
CD-15-00941	CLO_AA9728AA	2	2	2	0	1	76 (CC3)	0,1318
CD-15-00941	CLO_AA9603AA	3	3	3	0	1	76 (CC3)	0,023
CD-15-00941	CLO_BA0019AA	0	0	0	0	1	76 (CC3)	0,1343
CD-15-00941	CLO_AA9634AA	2	2	2	1	1	76 (CC3)	0,043
CD-15-00941	CLO_AA9808AA	0	0	0	0	1	76 (CC3)	0,1351
CD-15-00941	CLO_AA9698AA	1	1	2	1	1	76 (CC3)	0,1248

Die nahe genomische Verwandtschaft, die aufgrund der Analyse des Kerngenoms für die Isolate in den HC2 Clustern angenommen wurde, konnte für 59 % der paarweisen Vergleiche durch mindestens eine der vier berechneten SNP Distanzen bestätigt werden (Tabelle C.1). Eine Distanz von $\leq$2 SNPs wurde in 38 % der paarweisen Vergleiche für alle vier Referenzen erzielt. Bei 37 % der paarweisen Vergleiche wurde eine nahe genomische Verwandtschaft wiederum einheitlich ausgeschlossen. Bei der Verwendung einer Referenzsequenz, die dem gleichen CC wie die Genome in dem HC2 Cluster angehörte, wurden tendenziell höhere Differenzen im Vergleich zu den anderen drei beziehungsweise zwei verwendeten Referenzsequenzen berechnet (zum Beispiel HC2 Cluster 1 gegen Referenz M120 (CC1); Tabelle 3.7).
Die Unterschiede im akzessorischem Genom schienen die hohen SNP Distanzen zu bestätigen. Mit einer durchschnittlichen Jaccard Distanz von 0,1 unterschieden sich die Isolate in HC2 Cluster 1 deutlich im akzessorischem Genom und ließen eine nahe genomische Verwandtschaft nicht plausibel erscheinen. Des Weiteren lagen die Daten der Probennahme durchschnittlich 3,75 Jahre auseinander, was der durchschnittlichen Zeitdifferenz der 1.004 untersuchten Isolate in Kapitel 3.4.4 entspricht. Im Widerspruch dazu standen die für HC2 Cluster 76 erzielten Ergebnisse. Bei einer durchschnittlichen Jaccard Distanz von 0,1 im akzessorischem Genom wurden hier hauptsächlich SNP Distanzen von $\leq$2 berechnet (Tabelle 3.7). Besonders hervorzuheben ist hier, dass die nahe genomische Verwandtschaft bis auf einen paarweisen Vergleich von allen Referenzen bestätigt wurde. Insgesamt wurden für 17 % der paarweisen Vergleiche für mindestens eine Referenz eine SNP Distanz $\leq$2 und eine Jaccard Distanz von $\sim$ 0,05 berechnet (Tabelle C.1). Somit traten auch bei geringen SNP Distanzen hohe Differenzen im akzessorischen Genom auf.
In acht paarweisen Vergleichen wurde die nahe Verwandtschaft im Kerngenom durch identische akzessorische Genome bestätigt (Jaccard Distanz = 0; Tabelle C.1). In sechs Fällen lagen die berechneten SNP Distanzen allerdings über zwei. Wenn auch nur für einen geringen Anteil der paarweisen Vergleiche, so zeigten sich doch Unterschiede in den SNP und akzessorischen Distanzen, sodass die zur cgMLST-Analyse widersprüchlichen SNP Distanzen nicht immer durch Mutationen außerhalb des Kerngenoms erklärt werden konnten.
Ein weiterer zu beachtender Aspekt ist die Anzahl der nicht bestimmten Allele in den cgMLST Profilen der analysierten Genome. Da nicht bestimmte Allele paarweise von der Distanzberechnung in

der cgMLST-Analyse ausgeschlossen wurden, gibt es eine geringe Wahrscheinlichkeit, dass sich die zu vergleichenden Genome gerade in diesen ausgeschlossenen Loci unterscheiden. In der cgMLST-Analyse würden demnach mögliche Unterschiede nicht detektiert, die durch die SNP-Analyse erfasst werden. Ein Beispiel für solch einen Fall stellte HC2 Cluster 109 dar (Tabelle C.1), welches drei Isolate aus Kapitel 3.4.2 umfasste. Eine nahe genomische Verwandtschaft zu einem anderen Isolat in EnteroBase konnte nur für eines der drei Isolate durch die SNP-Analyse bestätigt werden. Die anderen Genome zeigten eine Distanz von bis zu 20 SNPs zu epidemiologisch nicht verwandten Genomen. Eine über dem Durchschnitt liegende Anzahl an nicht bestimmten Allelen von 39 beziehungsweise 6 in den cgMLST Allelprofilen der beiden Isolate schien die hohen SNP Distanzen zunächst zu erklären. Allerdings befanden sich die durch die SNP-Analyse detektierten Mutationen nur auf Loci, für die eine Allelnummer bestimmt werden konnte. Somit unterschieden sich die Genome nicht wie anfangs angenommen in den in der cgMLST-Analyse ausgeschlossenen Loci. Die SNPs befanden sich also außerhalb des Kerngenoms. Eine durchschnittliche Jaccard Distanz im akzessorischem Genom von 0,22 ließ den gleichen Schluss zu.

Bei der Analyse von, laut cgMLST-Analyse, nah verwandten Genomen resultierte die SNP-Analyse unter Verwendung unterschiedlicher Referenzen in abweichenden Distanzen. Neben der Anpassung der *seed*-Länge im *Mapping* Prozess erklärten sich die höheren SNP Distanzen im Vergleich zur cgMLST-Analyse auch durch die Detektion von Mutationen außerhalb des Kerngenoms. Es zeigte sich allerdings keine deutliche Tendenz, ob die Verwendung einer nah verwandten Referenzsequenz zu korrekteren Ergebnissen führte oder nicht. Dass neben der Wahl der Referenzsequenz noch weitere Aspekte in der SNP-Analyse zu beachten sind, wird im folgenden Kapitel gezeigt.

3.5.2 SNP-Analyse von hoch diversen Datensätzen

Wie schon in Kapitel 3.4.4 erwähnt, konnte für die auf Kerngenom-Allelunterschieden nah verwandten 1.004 Genome in EnteroBase eine genomische Distanz von ≤2 nur für 50 % der Einträge durch die SNP-EToKi-Analyse bestätigt werden. Die SNP-EToKi-Analyse verwendete die in EnteroBase zur Verfügung stehenden Assemblierungen und alignierte die *Contigs* der jeweiligen Assemblierungen an eine Referenzsequenz. Die Analyse wurde wie schon im vorangegangenen Kapitel mit vier Referenzsequenzen durchgeführt.

Tabelle 3.8: Anzahl der Isolate in paarweisen Vergleichen mit einer genomischen Distanz von 0, 1, 2 und *leq*2, ermittelt durch die cgMLST- und die EToKi-SNP Analyse für den ausgewählten Datensatz von 1004 Isolaten aus Kapitel 3.4.4.

Genomische Distanzen	Anzahl der Isolate in paarweisen Vergleichen (%)				
	Kerngenom Allelunterschiede	SNPs (R20291)	SNPs (CD196)	SNPs (CD630)	SNPs (M120)
0	159	76 (47,8 %)	76 (47,8 %)	67 (42,14 %)	67 (42,14 %)
1	371	157 (42,55 %)	155 (42,01 %)	154 (41,73 %)	136 (36,86 %)
2	716	319 (44,68 %)	321 (44,96 %)	317 (44,4 %)	285 (39,92 %)
≤2	1004	495 (49,5 %)	495 (49,5 %)	505 (50,5 %)	465 (46,5 %)

Interessanterweise resultierten die SNP-Analysen unter Verwendung der unterschiedlichen Referenzsequenzen in vergleichbaren Anteilen an Isolaten in paarweisen Vergleichen mit einer Distanz von ≤2 SNPs (46,5 % - 50,5 %; Tabelle 3.8). Trotz der Verwendung von *de novo* Assemblierungen war auch diese Analyse von der Wahl der Referenz abhängig, da die *Contigs*, wie die Illumina-*Reads* in der SNP-Analyse, an die Referenzsequenz aligniert wurden. Die voneinander abweichenden SNP Distanzen in der Analyse in Kapitel 3.5.1 konnten also nicht nur von der Wahl der Referenzsequenz abhängig sein. Um weitere Schritte in der SNP-Analyse definieren zu können, die eine Auswirkung auf die resultierenden SNP Distanzen haben, wurde für 816 der 1.004 Genome aus Kapitel 3.4.4 die *Read-Mapping* basierte SNP-Analyse durchgeführt (für die restlichen 188 standen keine Illumina-*Reads* zur Verfügung). Die Illumina-*Reads* wurden gegen die Referenz R20291

aligniert und die Konsensussequenzen wurden gebildet. Die folgenden Schritte wurden wie in Tabelle 3.9 beschrieben variiert.

Tabelle 3.9: Anzahl der paarweisen Vergleiche und der involvierten Isolate mit einer genomischen Distanz von ≤2. Die Distanzen wurden neben der cgMLST- und SNP-EToKi-Analyse mit verschiedenen Variationen der SNP-Analyse berechnet. Der analysierte Datensatz bestand aus 816 Genomen welche eine nahe Verwandtschaft von ≤2 Kerngenom-Allelunterschieden zu einem anderen Genom in EnteroBase, dessen zugehöriges Isolat in einem anderen Land isoliert wurde, aufzeigte (Kapitel 3.4.4; eigentlich 1.004 Genome, aber nur für 816 konnten die Illumina*Reads* heruntergeladen werden). ClonalFrameML und RecHMM wurden zur Detektion der Rekombinationen verwendet.

Analyse	Anzahl Vergleiche ≤2 (%)	Anzahl der Isolate (%)
cgMLST Analyse	1521 (100)	816 (100)
SNP-EtoKi-Analyse	392 (25,77)	373 (45,71)
SNP-Analyse (mit Rekombinationen, spaltenweiser Ausschluss von „N“ in der SNP Detektion)	1139 (74,88)	735 (90,07)
SNP-Analyse (mit Rekombinationen, paarweiser Ausschluss von „N“ in der SNP Detektion)	135 (8,88)	168 (20,59)
SNP-Analyse (ClonalFrameML, Detektion der SNPs basierend auf Rekombinationen korrigierten phylogenetischen Baum)	177 (11,64)	204 (25)
SNP-Analyse (ClonalFrameML, spaltenweiser Ausschluss von Rekombinationen und „N“ in der SNP Detektion)	1511 (99,34)	815 (99,88)
SNP-Analyse (ClonalFrameML, spaltenweiser Ausschluss von Rekombinationen und paarweisen Ausschluss von „N“ in der SNP Detektion)	763 (50,16)	555 (68,01)
SNP-Analyse (ClonalFrameML, paarweiser Ausschluss von Rekombinationen und „N“ in der SNP Detektion)	400 (26,3)	399 (48,9)
SNP-Analyse (RecHMM, paarweiser Ausschluss von Rekombinationen und „N“ in der SNP Detektion)	401 (26,36)	400 (49,02)

Wurden statt der Assemblierungen die Konsensussequenzen für die SNP-EToKi-Analyse verwendet, stieg der Anteil der Isolate, die eine genomische Verwandtschaft von ≤2 zu einem anderen Isolat in EnteroBase zeigten, um 3,3 % (Assembly-basiert: 45,71 %; *Mapping*-basiert: 49,02 %; Tabelle 3.9). Die höheren genomischen Unterschiede basierend auf den De-Novo-Assemblierungen könnten durch Fehler in dem Assemblierungsvorgang hervorgerufen worden sein.

Ein weiterer Unterschied zwischen der SNP- und der SNP-EToKi-Analyse war die Detektion und der Ausschluss von rekombinanten Elementen in den zu vergleichenden Genomen. Die Wahl des bioinformatischen Werkzeuges zur Detektion der rekombinanten Stellen hatte keinen Einfluss auf das Ergebnis (RecHMM: 49,02 %; ClonalFrameML: 48,9 % an Isolaten in paarweisen Vergleichen ≤2 SNPs; Tabelle 3.9). Allerdings war die Art des Ausschlusses der detektierten Stellen entscheidend für die Ergebnisse. Die SNP-EToKi-Analyse extrahierte die Stellen mithilfe des Werkzeuges RecFilter, welches die Stellen paarweise aus dem Alignment entfernt. Die rekombinanten Stellen die die Software ClonalFrameML ermittelte, wurden bis zu diesem Zeitpunkt immer spaltenweise aus dem Alignment entfernt. Da bis zu dieser Analyse nur nah verwandte Datensätze untersucht wurden, stellte dies kein Problem dar. Rekombinationen

traten allerdings unterschiedlich oft in verschiedenen *Lineages* von *C. difficile* auf, sodass sich die betroffenen Stellen voneinander unterschieden. Bei einem hoch diversen Datensatz ging somit durch spaltenweisem Ausschluss ein Großteil der genomischen Information verloren (spaltenweiser Ausschluss: 99,9 %; paarweiser Ausschluss: 48,9 % an Isolaten in paarweisen Vergleichen ≤2 SNPs; Tabelle 3.9). Bei der Berechnung der SNPs anhand der Astlängen des durch die ClonalFrameML-Analyse korrigierten phylogenetischen Baumes detektierte die SNP-Analyse weniger paarweise Vergleiche mit ≤2 SNPs als bei einem paarweisen Ausschluss der rekombinanten Elemente (Umrechnung der Astlängen: 25 %; paarweiser Ausschluss: 48,9 %; Tabelle 3.9). Dies ist zum einen darin begründet, dass bei dem paarweisem Ausschluss absolute Distanzen berechnet, bei der Umrechnung des Baumes aber relative Distanzen berechnet wurden, da hier die Anzahl der Merkmale durch die Filterung der Rekombinationen für jedes Genom unterschiedlich ist. Zum anderen wurden bei der Berechnung der Astlängen des rekombinanten-korrigierten phylogenetischen Baumes während der ClonalFrameML-Analyse die erwarteten Mutationen des der Analyse zugrunde liegenden Maximum-Likelihood Baumes verwendet. Diese erwarteten Mutation können nach Filterung der Rekombinanten von den tatsächlich vorliegenden Mutationen abweichen und somit zu größeren Distanzen zwischen den Genomen führen.
Ein weiterer Aspekt war die Bestimmung der SNP Distanzen aus dem rekombinanten-korrigierten Alignment und der Berücksichtigung der Basen, die eine unzureichende Qualität aufwiesen um in der Konsensussequenz bestimmt werden zu können („N“). Auch hier war der spaltenweise Ausschluss bei der SNP Detektion in dem Alignment bis zur Analyse eines hoch diversen Datensatzes kein Problem. Bei hoch diversen Sequenzen, die gegen die gleiche Referenzsequenz aligniert wurden, fanden sich Regionen mit unbestimmten Basen über die ganze resultierende Konsensussequenzen verteilt. Schnitt man jede Spalte, welche ein „N“ beinhaltete aus dem Alignment raus, verlor man einen Großteil der genomischen Information, was in fast identischen paarweisen Vergleichen resultierte (99,9 %; Tabelle 3.9).
Letztendlich wurde durch die Analyse deutlich, dass der Ausschluss von Rekombinationen besonders bei der Analyse von nah verwandten Isolaten von großer Bedeutung ist. Rekombinationen riefen zwischen nah verwandten Isolaten große genomische Distanzen hervor, da sie meist eine Vielzahl an Mutationen in das Genom einbringen. Somit konnte die nahe genomische Verwandtschaft nicht durch die SNP-Analyse detektiert werden (mit Rekombinationen: 20,59 %; ohne Rekombinationen: 48,9 % an Isolaten in paarweisen Vergleichen ≤2 SNPs).

Das abschließende Kapitel des Ergebnisteils verdeutlicht, was für eine Auswirkung die kleinsten methodischen Schritte in den Analysen auf die biologischen Schlussfolgerungen haben können. Von einer Einstellung im *Read-Mapping* über die Wahl der Referenz bis hin zur Handhabung von nicht bestimmten Basen beziehungsweise Allelen haben nahezu alle Schritte der cgMLST- und SNP-Analyse die Anteile an Isolaten mit naher genomischer Verwandtschaft zu einem anderen Isolat beeinflusst. Die Analysen sind sensibel im Bezug auf den zu untersuchenden Datensatz und bestimmte Schritte müssen bei der Analyse von hoch diversen Datensätzen angepasst werden. Demnach sollten sowohl bei der Anwendung von cgMLST- als auch von SNP-Analysen die hier aufgezeigten Aspekte beachtet werden, um die Detektion von nahen genomischen Verwandtschaften so korrekt wie nur möglich durchzuführen.

Kapitel 4

Diskussion

Das Vorkommen von Bakterien die nosokomiale Infektionen verursachen ist weltweit ein großes Problem, das sowohl die Gesundheit vieler Patienten gefährdet als auch das Gesundheitssystem finanziell belastet. Um einer weiteren Ausbreitung der Bakterien effektiv entgegenwirken zu können, sollten nah verwandte pathogene Stämme innerhalb und zwischen Krankenhäusern, wie auch auf internationaler Ebene schnell detektiert werden, um so mögliche Transmissionsrouten zu erkennen. Für entsprechende genomische Analysen sind eine Vielzahl an bioinformatischen Werkzeugen verfügbar, die meist keine graphische Oberfläche besitzen und über die Kommandozeile bedient werden müssen. Zudem können veränderte Einstellungsparameter zu unterschiedlichen Ergebnissen führen, was ein detailliertes Verständnis der angewendeten Methoden erfordert. Dadurch sind genomische Analysen für Menschen ohne bioinformatischem Hintergrund oft nicht durchführbar.

Erste Alternativen zu Kommandozeilen-basierten Methoden bieten die kommerziell verfügbaren Softwares SeqSphere$^+$ und BioNumerics, die sich über graphische Oberflächen bedienen lassen. Sequenzdaten können anhand von implementierten bioinformatischen Werkzeugen analysiert und genomische Beziehungen durch verschiedene Algorithmen dargestellt werden. Die Analysen begrenzen sich jedoch auf den eigenen Datensatz.

Mit der Entwicklung von EnteroBase und den dort implementierten bioinformatischen Werkzeugen bietet sich eine neue Möglichkeit für Epidemiologen, Zusammenhänge zwischen Pathogenen auf Basis ihrer Genome herzustellen und einen globalen Überblick über deren Vorkommen zu erhalten. Anhand der verfügbaren *C. difficile* Datenbank können eigene Datensätze mit einer stetig wachsenden Sammlung an *C. difficile* Genomen und den zugehörigen Metadaten der Isolate verglichen werden. Zudem ist es möglich phylogenetische Bäume für eine Vielzahl an Genomen zu berechnen, ohne das ein bioinformatischer Hintergrund notwendig ist.

Die Analyse der umfangreichen Datenbank erlaubte einen noch nie erfassten Einblick in die Populationsstruktur von *C. difficile* und ermöglichte es durch die implementierte hierarchische Clusterung neue Typisierungsmethoden anzuwenden sowie Ausbruchsdetektionen vorzunehmen. Zunächst galt es allerdings die bioinformatischen Werkzeuge in EnteroBase mit standardmäßig in der Literatur verwendeten Methoden zu vergleichen. Anhand dessen sollte demonstriert werden, dass EnteroBase für molekularepidemiologische Analysen anwendbar ist, sowohl für pandemische und endemische Untersuchungen als auch für Ausbruchsdetektionen.

4.1 Bewertung und Vergleich der bioinformatischen Methoden in der genomischen Epidemiologie

4.1.1 Kritische Betrachtung der verfügbaren Methoden

Vergleich der verfügbaren cgMLST Schemata

Die Ergebnisse der vorliegenden Arbeit basieren hauptsächlich auf dem frei verfügbaren cgMLST Schema in EnteroBase, das auf Grundlage von 442 *C. difficile* Genomen entwickelt wurde. Durch die sorgfältige Auswahl der Referenzgenome umfasst das cgMLST Schema Genomabschnitte, die zu 99,96 % in den 13.515 untersuchten Genomen vorkamen und von denen 99,94 % einer Allelnummer zugeordnet werden konnten. Zusätzlich deutete ein Diskriminierungsindex von 0,99 der hierarchischen Cluster HC0 an, dass mit Hilfe des cgMLST Schemas ein Großteil der Genome in individuelle Cluster sortiert werden konnte. Es kam jedoch vor, dass multiple Genome in einem HC0 Cluster zusammengefasst wurden (siehe Kapitel 3.1, Seite 29). Dies ist zum Beispiel auf Studien zurückzuführen, die die Stabilität eines *C. difficile* Stammes untersuchten und diesen dafür mehrmals sequenzierten (zum Beispiel *BioProject* Nummer PRJEB2195). Des Weiteren liegen etliche Studien vor, die identische beziehungsweise nah verwandte Genome beinhalten [153], [155], [156], sodass ein paar der Genome wahrscheinlich in ein HC0 Cluster fallen.
Für das cgMLST Schema in BioNumerics waren keine Werte für das Auflösungsvermögen oder für die Typisierbarkeit verfügbar. Die geringere Anzahl an cgMLST Loci in dem Schema lässt allerdings eine geringere Auflösung im Vergleich zu dem Schema in EnteroBase vermuten (BioNumerics: 1.999; EnteroBase: 2.556 Loci). Das gleiche gilt auch für das cgMLST Schema in SeqSphere$^+$ (2.270 Loci), für das in der zugehörigen Publikation keine Analyse des Auflösungsvermögens erwähnt wird[67]. Die Typisierbarkeit des Schemas wurde wiederum anhand dreier Datensätze (Genome zweier Ausbrüche, 70 Genome die die Toxinotypen von *C. difficile* repräsentieren und 2.268 assemblierte Genome, die auf NCBI zur Verfügung standen) bestimmt[67]. Mit Anteilen an durchschnittlich bestimmten Allelen zwischen 99,1 - 99,5 % erzielte das SeqSphere$^+$ Schema eine ähnlich hohe Typisierbarkeit wie das cgMLST Schema in EnteroBase.
Dadurch, dass EnteroBase sowohl manuelles Hochladen von *Read*-Daten erlaubt, als auch alle verfügbaren Illumina-*Reads* aus den *Short-Read-Archives* mit in die Datenbank einbringt, besteht die Gefahr von doppelten Einträgen. Würde ein Nutzer seine Daten mit EnteroBase analysieren und diese später in einem der Repositorien veröffentlichen, würde EnteroBase diese erneut herunterladen und analysieren. Des Weiteren werden auch Genome mit in die Datenbank aufgenommen, die nicht aus humanen, tierischen oder Umweltquellen stammen, sondern zum Beispiel im Labor aus einem Darmmodell isoliert wurden[157]. Die Datenbank erfordert demnach eine regelmäßige Kuratierung, sodass doppelte Einträge vermieden und identische Stämme als ein so genannter „*Uberstrain*“ zusammengefasst werden können.
Die cgMLST-Analyse in EnteroBase ermöglicht den Vergleich der eigenen Genome mit allen Einträgen in der Datenbank. So können genomische Vergleiche nicht nur zwischen ausbruchsverwandten Isolaten erfolgen, sondern auch zwischen Isolaten, die nicht im Rahmen des zu untersuchenden Ausbruchs isoliert wurden. Dadurch können Ausbruchsanalysen verbessert und nicht offensichtliche Zusammenhänge aufgedeckt werden [45]. Durch das Hochladen der eigenen Daten auf EnteroBase stehen diese allerdings jedem anderen Nutzer der Datenbank zur Verfügung. Sensible Metadaten der Isolate können jedoch in einem so genannten „*Custom View*“ abgespeichert und dadurch privat gehalten werden.
EnteroBase lässt eine einheitliche Analyse aller Genome in der Datenbank zu, indem nur unbehandelte Sequenzdaten (*Short-Reads*) auf die Plattform hochgeladen werden können, die durch eine standardisierte Pipeline assembliert und analysiert werden. Das Hochladen von bereits erstellten Assemblierungen ist nur nach Absprache mit den Entwicklern möglich. Des Weiteren werden durch die Verwendung des Werkzeuges Pilon Misassemblierungen vermieden[68]. Die Qualität der resultierenden Assemblierungen wird zusätzlich anhand der Metriken in Tabelle 2.1 (Seite 15) überprüft, bevor die Daten der Datenbank hinzugefügt werden. Während SeqSphere$^+$ die *fastq*-Dateien der hochgeladenen Illumina-*Reads* vor der Assemblierung auf ihre Qualität prüft und die Assemblierungen mit einer Abdeckung ≤ 50 von weiteren Analysen ausschließt[67],

[158], lagen für BioNumerics keine weiteren Informationen dazu vor. Bei beiden Softwares können neben *Short-Reads* auch selbst erstellte Assemblierungen für die cgMLST Typisierung verwendet werden. Auch diese müssen, zumindest bei SeqSphere$^+$, erst den Qualitätscheck bestehen. Die Überprüfung der Qualität verspricht jedoch nicht eine einheitliche Analyse der Daten. Frühere Studien weisen darauf hin, dass die Verwendung unterschiedlicher Assemblierungs-Werkzeuge sowie die dabei gewählten Einstellungen Einfluss auf die resultierende Assemblierung haben und dadurch unterschiedliche genomische Zusammenhänge zwischen zwei Isolaten geschlussfolgert werden können[64], [159]. Demnach besteht die Gefahr, dass die cgMLST-Analyse selbst erstellter Assemblierungen genomische Distanzen erzeugt, die auf die Verwendung unterschiedlicher Werkzeuge zurückzuführen sind und keine tatsächlichen Genomunterschiede repräsentieren. Genomische Zusammenhänge zwischen Isolaten können durch die hierarchische Clusterung der Genome in EnteroBase unkompliziert auf verschiedenen Distanzebenen untersucht werden. Ohne zusätzliche Analyseschritte ermöglichen die hierarchischen Cluster identische Isolate (HC0), mögliche Transmissionsketten (HC2), Pandemien (HC10) und endemische Stämme (HC150) zu identifizieren. Eine Distanzmatrix, die die paarweisen Kerngenom-Allelunterschiede der Genome beinhaltet, kann in EnteroBase allerdings nicht erstellt werden. In der vorliegenden Arbeit wurden die Allelprofile hierfür in BioNumerics importiert, um dort eine Distanzmatrix zu berechnen (siehe Kapitel 2.2.2, Seite 19). Paarweise Distanzen sind in EnteroBase anhand der Minimum-Spanning-Bäume erkennbar, bei denen die Astlänge die Anzahl der Kerngenom-Allelunterschiede widerspiegelt. Auch in SeqSphere$^+$ wird keine exportierbare Distanzmatrix erstellt. Es sind jedoch mehr Informationen verfügbar, wie zum Beispiel die Anzahl der Kerngenom-Allelunterschiede zu dem nächsten verwandten Genom. Des Weiteren gibt SeqSphere$^+$ noch weitere nützliche Parameter für jedes analysierte Genom an, darunter zum Beispiel die Anzahl an nicht bestimmbaren Allelen und die Qualität der Allelsequenzen. Diese Angaben könnten die Fehlersuche bei überraschend resultierenden genomischen Zusammenhängen zwischen zwei Isolaten erleichtern.
Andere Studien weisen darauf hin, dass Ganzgenomanalysen von Nicht-Bioinformatikern am besten mit Hilfe kommerzieller Software durchgeführt werden sollten[160]. Die vorliegende Arbeit zeigt, dass EnteroBase zumindest für Analysen der dort enthaltenen Bakterien eine frei verfügbare Alternative zu den kostenpflichtigen Softwares bietet. Zudem ermöglicht EnteroBase den Vergleich der eigenen Daten mit einer umfangreichen Datenbank, die sämtliche verfügbaren Genomsequenzen von öffentlichen Repositorien (NCBI, EBI) von *C. difficile* umfasst.

Die cgMLST Schemata wurden entwickelt, um eine effektivere und schnellere Alternative für die in der molekularen Epidemiologie als Goldstandard bezeichnete SNP-Analyse zu finden [68], [161]. Neben der, besonders bei umfangreichen Datensätzen langen Rechenzeit, sind die verschiedenen Werkzeuge für die SNP-Analyse aufgrund der Vielzahl an Einstellungsmöglichkeiten und der Kommandozeilen-basierten Bedienung oft für Nicht-Bioinformatiker nur schwer anwendbar. Die vorliegende Arbeit zeigte weitere Schwierigkeiten der SNP-Analyse auf, die oft nicht im Detail erwähnt und zudem selten berücksichtigt werden.

Die Fallstricke der SNP-Analyse

Zu Beginn der *Read-Mapping* basierten SNP-Analyse steht die Wahl der Referenzsequenz, gegen die die *Reads* aligniert werden sollen. Die Wahl der Referenzsequenz wirkte sich sowohl bei der Analyse von nah verwandten als auch bei hoch diversen Datensätzen auf die resultierenden SNP Distanzen aus, was insbesondere bei der Detektion von Transmissionswegen zu falschen Schlüssen führen könnte. Das *Mapping* der Illumina-*Reads* nah miteinander verwandter Isolate gegen ein geschlossenes Genom aus dem gleichen HC150 Cluster resultierte tendenziell zu höheren paarweisen SNP Distanzen als gegen genomisch distinktere Referenzgenome (Tabelle 3.7, Seite 52). Dies ist darauf zurückzuführen, dass die erstellte Konsensussequenz nur Regionen beinhaltet, die sich auch in der Referenzsequenz befinden. Sequenzabschnitte, die nur in den Illumina-*Reads* zu finden sind, können nicht an die Referenzsequenz aligniert werden und gehen somit

verloren[154], [162].

Auch in der Literatur wurde schon darauf hingewiesen, dass Illumina-*Reads* im besten Fall gegen eine nah verwandte Referenzsequenz aligniert werden sollten[76], [159]. Dies gestaltet sich allerdings als schwierig, wenn keine weiteren Typisierungsinformationen für die Isolate der zu rekonstruierenden Sequenzen vorliegen. Hier bietet die umfangreiche Datenbank in EnteroBase einen Vorteil: Die automatische Zuordnung der entstandenen Assemblierungen zu einem HC150 Cluster erleichtert die Suche nach einem passendem, nah verwandtem Referenzgenom. Befindet sich kein geeignetes Referenzgenom in dem gleichen HC150 Cluster, kann ein phylogenetischer Baum, der alle verfügbaren geschlossenen Genome und die zu analysierenden Genome beinhaltet, berechnet werden, um so den am nächsten verwandten Eintrag zu ermitteln. Bei der Analyse eines diversen Datensatzes, bei dem die Genome in multiple HC150 Cluster fallen, könnte die Wahl auf ein Referenzgenom fallen, das eine gleiche genomische Distanz zu allen Clustern aufzeigt. Ein ähnlicher Ansatz wurde auch schon von Bush et al. vorgeschlagen[76]. Eine weitere Herangehensweise verwendet ein Datensatz-internes Genom als Referenz. Dafür müsste ein Set an Illumina-*Reads* assembliert werden, sodass die anderen *Reads* gegen die entstandene Assemblierung gemappt werden können[162]. Das initiale Assemblieren aller zu untersuchenden Illumina-*Reads* ist die Grundlage für einen weiteren Ansatz, bei dem die Assemblierungen geclustert werden und die Wahl einer passenden Referenz für jedes Cluster separat erfolgt[72], [163]. Die weiteren Schritte der SNP-Analyse werden dann für jedes Cluster einzeln durchgeführt. Diese Methode ist zur korrekten Detektion von nahen genomischen Verwandtschaften nützlich, gewährt aber keinen Einblick über die genomischen Distanzen zwischen den Isolaten aus unterschiedlichen Clustern.

Der Ausschluss von Rekombinationen und nicht bestimmbaren Basen stellte sich im weiteren Verlauf der SNP-Analyse als zusätzlicher Einflussfaktor heraus. Der paarweise beziehungsweise spaltenweise Ausschluss der betroffenen Stellen wirkte sich stark auf den Anteil an paarweisen Vergleichen mit einer genomischen Distanz von $\leq$2 SNPs aus (Tabelle 3.9, Seite 54). Während bei diesen Ansätzen die SNP Distanzen anhand der Punktmutationen in dem SNP-Alignment ermittelt werden, wird in vielen Publikationen eine andere Vorgehensweise gewählt. Dabei werden die genomischen Distanzen basierend auf den Astlängen des rekombinationen-korrigierten phylogenetischen Baumes berechnet. Auch diese Methode führte zu abweichenden Anteilen an paarweisen Vergleichen mit $\leq$2 SNP Distanzen (Tabelle 3.9, Seite 3.9). Diese Methode wird oft verwendet, da sie die Evolution der analysierten Genome mit berücksichtigt. Die Astlängen, und somit die berechneten genomischen Distanzen, eines phylogenetischen Baumes werden allerdings durch das verwendete Modell zur Berechnung des Baumes beeinflusst. Das hier verwendete Maximum-Likelihood-Modell kann zwar bei komplexen Datensätzen die Astlängen der längeren Äste akkurater bestimmen, zeigt aber dafür Defizite bei der Bestimmung der kürzeren Astlängen[162], [164]. Der hier verwendete rekombinationen-korrigierte phylogenetische Baum wurde durch ClonalFrameML ausgegeben. ClonalFrameML lässt neben anderen Einflussfaktoren auch die durchschnittliche Astlänge des Ausgangsbaumes mit in die Kalkulation der Astlängen des korrigierten Baumes einfließen[74]. Da die durchschnittliche Astlänge bei hoch diversen Datensätzen als recht hoch anzunehmen ist, könnten dadurch ursprünglich kürzere Astlängen überschätzt werden. Weiterhin wurde gezeigt, dass das Verhältnis zwischen Transitionen und Transversionen mit den Astlängen eines Maximum-Likelihood-Baumes korreliert[164]. Dieses Verhältnis wird in ClonalFrameML-Analysen durch den Parameter *kappa* mit in die Analyse einbezogen. Für die Analysen der vorliegenden Arbeit wurde dieser Wert nicht extra berechnet, sondern ein Wert von 2 verwendet, da dieses Verhältnis für homologe DNA-Stränge angenommen wird. Ob und wie sich das Verhältnis zwischen Transitionen und Transversionen auf die Astlängen auswirkt müsste allerdings noch näher untersucht werden. Die Bestimmung der SNP Distanzen wird in der Literatur sowohl über das Alignment als auch über die Astlängen eines phylogenetischen Baumes durchgeführt. Die in dieser Arbeit erzielten Ergebnisse deuten allerdings an, dass SNP Distanzen eines hoch diversen Datensatzes durch das Umrechnen der Astlängen eines phylogenetischen Baumes überschätzt werden könnten und für solche Datensätze die SNP Distanzen anhand des Alignments berechnet werden sollten.

Der Ausschluss von nicht bestimmten Basen in der Konsensussequenz gegenüber der Referenz sollte besonders bei einem diversen Datensatz paarweise erfolgen, da sich die zu untersuchenden Sequenzen stark voneinander

unterscheiden können und an verschiedenen Stellen betroffen sind. Dadurch würde bei einem spaltenweisen Ausschluss dieser Stellen ein Großteil der genomischen Information verloren gehen.

Die Analysen der vorliegenden Arbeit bestätigen die zum Teil zuvor diskutierten Probleme der SNP-Analyse und bekräftigen die Schwierigkeit der Standardisierung dieser Methode durch weitere Argumente. Zudem wird deutlich, dass für die SNP-Analyse sowohl ein gewisses bioinformatisches Verständnis, als auch die Fähigkeit der Anwendung komplexer Werkzeuge ohne graphische Oberfläche von Nöten ist. Die unterschiedliche Handhabung eines Datensatzes oder die Wahl verschiedener Parameter bei der Anwendung der bioinformatischen Werkzeuge kann zu unterschiedlichen genomischen Distanzen und dadurch zu abweichenden epidemiologischen Schlussfolgerungen führen.

4.1.2 Eignung der cgMLST für die molekulare Epidemiologie

Die vorliegende Arbeit präsentiert erstmals einen umfassenden Vergleich zwischen cgMLST- und SNP-Analyse für *C. difficile* Genome und zeigt die analytischen Herausforderungen der genomischen Epidemiologie auf. Die cgMLST-Analyse mit EnteroBase ergab vergleichbare Ergebnisse wie die SNP-Analyse und konnte so als alternative Methode zur Detektion von Transmissionswegen verwendet werden.

Wie im vorangegangenen Kapitel diskutiert, beinhaltet die SNP-Analyse besonders für nicht erfahrene Bioinformatiker herausfordernde Aspekte. Trotzdem werden mögliche Transmissionswege pathogener Bakterien meist anhand von SNP Distanzen untersucht [72], [165]. So hat sich für *C. difficile* ein Grenzwert von $\leq$2 SNPs zur Detektion von Isolaten etabliert, die mit einer 95 %igen Wahrscheinlichkeit einer Transmissionskette angehören[72]. Wurden Ausbrüche mit der cgMLST-Analyse untersucht, so stimmten die Werte der Kerngenom-Allelunterschiede weitestgehend mit den SNP Distanzen überein (Abbildung 3.7, Seite 37). Bei einem Wert von $\leq$2 Kerngenom-Allelunterschieden betrugen mit einer 88 %igen Wahrscheinlichkeit die entsprechenden SNP Distanzen ebenfalls einen Wert von $\leq$2 (Abbildung 3.8, Seite 38). Auch Bletz et al. konnten anhand der Reanalyse zweier publizierter Ausbrüche diesen Grenzwert für direkte Transmissionswege mit der cgMLST-Analyse in SeqSphere$^+$ nachbilden[67]. Zur Detektion von Isolaten, die wahrscheinlich einem Ausbruchs-Klon angehören, definierten sie allerdings einen Grenzwert von $\leq$6 Kerngenom-Allelunterschieden[67]. Dieser Wert scheint recht weit gefasst, besonders wenn man berücksichtigt, dass mittels des SeqSphere$^+$ cgMLST Schemas im Rahmen der vorliegenden Arbeit vergleichsweise geringere Distanzen berechnet wurden als mit der cgMLST-Analyse in EnteroBase, was wiederum auf die niedrigere Anzahl an cgMLST Loci zurückzuführen ist (SeqSphere$^+$:2.270 Loci, EnteroBase: 2.556 Loci). Die geringeren Distanzen wie auch der hohe Grenzwert von $\leq$6 Kerngenom-Allelunterschieden birgen demnach die Gefahr, dass durch die SeqSphere$^+$ cgMLST-Analyse falsch positive Isolatepaare einer Transmissionskette zugeordnet werden.

Quantitative Diskrepanzen zwischen SNP Distanzen und Kerngenom-Allelunterschieden werden zudem von Phänomenen verursacht, die unabhängig von den verwendeten cgMLST Schemata sind. Höhere SNP Distanzen können zum Beispiel darauf zurückzuführen sein, dass multiple Punktmutationen in einer Allelsequenz vorkommen. Die cgMLST-Analyse würde dies als nur einen Unterschied verzeichnen. Des Weiteren können SNPs auch auf Genomabschnitten auftreten, die nicht durch die Loci in den cgMLST Schemata erfasst werden (siehe auch Tabelle 3.6, Seite 50). Resultiert die cgMLST-Analyse in höheren Genomunterschieden als die SNP-Analyse wird dies meist durch Insertionen und/oder Deletionen in den Genomen verursacht[69]. Diese werden in der SNP-Analyse nicht berücksichtigt.

Trotz der hier diskutierten Unstimmigkeiten zu der SNP-Analyse wurden durch die cgMLST-Analyse im weiteren Verlauf der Arbeit erfolgreich Transmissionswege aufgedeckt, sowohl zwischen klinischen Isolaten als auch bei Umweltproben (Kapitel 3.4.2 und 3.4.3, diskutiert in Kapitel 4.2.2).

Bei Datensätzen, die epidemiologisch nicht zusammenhängende Isolate beinhalteten, wichen die durch die cgMLST- und SNP-Analyse berechneten genomischen Distanzen allerdings deutlich voneinander ab. Es wurden weitaus mehr nahe Verwandtschaften durch die cgMLST- als durch die SNP-Analyse ermittelt

(Kapitel 3.4.4, Seite 44). Hier schien ein Großteil der Punktmutationen außerhalb der cgMLST Loci zu liegen, was zu einer näheren Analyse des akzessorischen Genoms motivierte.

Die Analyse des akzessorischen Genoms

Die Ergebnisse der vorliegenden Arbeit demonstrieren, dass unabhängig voneinander isolierte Isolate trotz eines fast identischen Kerngenoms deutliche Unterschiede im akzessorischem Gengehalt aufweisen können. Um sicherzustellen, dass diese Unterschiede keine Artefakte des wgMLST Schemas in EnteroBase sind, wurden die Daten zusätzlich mit der *k-mer* basierten PopPUNK-Analyse untersucht. Die Analysen führten zu vergleichbaren Ergebnissen, jedoch resultierte die PopPUNK-Analyse des vorausgewählten Datensatzes in eine leicht höhere Anzahl an paarweisen Vergleichen mit identischem akzessorischem Gengehalt als die Analyse des wgMLST-Schemas (Tabelle 3.5, Seite 47).
Die Diskrepanzen zwischen den beiden angewendeten Methoden könnten dadurch bedingt sein, dass PopPUNK die Genome auf eine so genannte *Sketch* reduziert[141]. In dieser Arbeit wurden die *C. difficile* Genome beispielsweise durch die Wahl einer *Sketch*-Größe von 10^5 und der maximalen *k-mer* Länge von 29 Basen in höchstens 69 % ihrer Genomlänge miteinander verglichen (2.9 Megabasen). Im Gegensatz dazu besteht das wgMLST Schema in EnteroBase aus 11.490 Loci, die alle kodierenden Bereiche der 442 Referenzgenome umfassen. Demnach ist es wahrscheinlich, dass das wgMLST Schema in EnteroBase Genomabschnitte berücksichtigt, die bei der PopPUNK-Analyse nicht erfasst werden und dadurch zu leicht niedrigeren Distanzen im akzessorischen Gengehalt führt.
Des Weiteren wird in PopPUNK die An- und Abwesenheit akzessorischer Genomabschnitte zwischen zwei Isolaten basierend auf der kleinsten verwendeten *k-mer* Länge von 17 Basen verglichen. Die Loci im wgMLST Schema wiederum haben eine durchschnittliche Länge von 854 Basen. Eine Deletion über solch eine Länge an der gleichen Stelle in zwei Genomen ist um einiges unwahrscheinlicher wie wenn es sich um eine 17 Basen lange Deletion handelt. Folglich tendiert die PopPUNK-Analyse eher dazu in paarweisen Vergleichen ein übereinstimmendes akzessorisches Genom zu detektieren als die Analyse des wgMLST Schemas.

Vor- und Nachteile der verwendeten Methoden und weiterer Anwendungen

Neben der Einträge der *C. difficile* Datenbank in EnteroBase ist zusätzlich die Kuratierung der Allelsequenzen aller MLST-Loci notwendig, sodass eine qualitativ hochwertige MLST-Typisierung gewährleistet werden kann. *K-mer* basierte Methoden wie PopPUNK oder hash-cgMLST bieten eine datenbankunabhängige Alternative zu der cgMLST-Analyse um genomische Beziehungen zwischen Isolaten zu untersuchen[64], [141]. Bei der hash-cgMLST Anwendung wird zum Beispiel anhand der Loci des cgMLST Schemas in SeqSphere$^+$ ein lokales cgMLST Allelprofil für jedes Genom erstellt. Um die Analyse serverunabhängig durchführen zu können, werden die Allelsequenzen für jeden Locus in individuelle *Hashes* überführt. Laut den Entwicklern der Anwendung werden dadurch reproduzierbare „hash-cgMLST Allelprofile“ erstellt, deren paarweise Distanzen zu vergleichbaren Werten wie die in SeqSphere$^+$ erstellten cgMLST Allelprofile führen[64].
Im Gegensatz zu den *k-mer* basierten Methoden stellt sowohl das cgMLST als auch das wgMLST Schema einen festgelegten Rahmen dar, in denen Genome miteinander verglichen werden. Besonders bei Bakterien wie *C. difficile*, die ein offenes Pangenom besitzen, werden neu eingebrachte Gene nicht in der Analyse berücksichtigt[114]. Durch die Anwendung PopPUNK können im Gegensatz dazu Kern- und akzessorisches Genom flexibel bestimmt werden[141]. Dies könnte sich besonders bei der Analyse von neu aufkommenden endemischen Stämmen von *C. difficile* als praktisch erweisen, da diese unter Umständen neue, im wgMLST Schema nicht repräsentierte, Genomabschnitte in sich tragen.
Ein weiterer zu berücksichtigender Aspekt, der alle bisher erwähnten Methoden betrifft, ist die Wahl der *Read-Mapping*- oder assemblierungsbasierten Detektion von nahen genomischen Verwandtschaften. Die *Mapping*-basierte SNP-Analyse resultierte für eine leicht höhere Anzahl an Isolaten in einer genomischen Distanz von $\leq$2, als die assemblierungsbasierte SNP-Analyse (Tabelle 3.9, Seite 54). Wie auch schon

in der Publikation von Eyre et al. beschrieben, können durch die assemblierungsbasierte SNP-Analyse falsch-negative Distanzen berechnet werden (>2 SNPs)[64]. Dieses Verhalten wurde auf Misassemblierungen zurückgeführt, die eine typische Begleiterscheinung des Assemblierungsprozesses sind[64]. Assemblierungen, die durch die Pipeline in EnteroBase erstellt wurden, sollten allerdings durch die zusätzliche Korrektur durch das Programm Pilon weitestgehend von falsch assemblierten Abschnitten befreit sein. Erneut wird hier deutlich, dass Illumina-*Reads* einheitlich prozessiert und entstandene Assemblierungen qualitätskontrolliert werden sollten, um reproduzierbare Ergebnisse erzeugen zu können.
Des Weiteren wurde in dieser Arbeit gezeigt, dass auch durch die *Mapping*-basierte SNP-Analyse falsch-positive beziehungsweise falsch-negative Distanzen detektiert werden können. Dadurch, dass sich eine Vielzahl an kurzen *Read*-Fragmenten an dieselbe Stelle des Referenzgenoms alignierten, wurden für das betroffene Genom fälschlicherweise Punktmutationen gegenüber der anderen Genome angenommen (Abbildung 3.16, Seite 51). Das Alignieren von kurzen *Read*-Fragmenten wurde durch die Wahl des BWA-MEM-Algorithmus für das *Read-Mapping* ermöglicht, da hier *Reads*, die mit unterschiedlichen Stellen des Referenzgenoms übereinstimmen, aufgesplittet werden können. Korrigiert werden konnte dies durch die Wahl einer längeren *seed*-Länge. Je länger man allerdings die *seed*-Länge wählt, desto unwahrscheinlicher wird eine komplette Übereinstimmung der *Reads* über die gewählte Länge. Nicht passende *Reads* werden nicht aligniert und die enthaltene genomische Information geht verloren[166]. Die ideale *seed*-Länge könnte sich aus den evolutionären Abständen und der Ähnlichkeit zwischen den untersuchten Genomen ergeben, würde somit aber wieder Voruntersuchungen des Datensatzes benötigen[166]. Besonders bei der Analyse von epidemiologisch zusammenhängenden Isolaten ist für das *Read-Mapping* gegen eine nah verwandte Referenzsequenz eine längere *seed*-Länge zu empfehlen (zum Beispiel wie hier angewendet eine Länge von 30 Basen). Ähnliche Genome sollten über eine längere Einheit an Basen übereinstimmen, sodass das Problem von kurzen, falsch gemappten *Read*-Fragmenten umgangen werden kann. Durch die Einstellungsmöglichkeiten im *Read-Mapping*, sowie die Qualitätsfilterung der Punktmutationen während des *Variant-callings*, lassen sich falsch bestimmte Mutationen in dem *Mapping*-basiertem Ansatz besser kontrollieren als in der assemblierungsbasierten SNP- und cgMLST-Analyse[64]. Falsch detektierte SNPs könnten durch stringenteres filtern der Qualität der Assemblierungen vermieden werden. In EnteroBase wird dies durch Pilon und der anschließenden Qualitätsfilterung der Assemblierungen schon umgesetzt.

Eine gute Typisierungsmethode sollte reproduzierbar und unabhängig von Nutzer, Zeit und Ort durchführbar sein[45]. Die Reproduzierbarkeit stellt bei der SNP-Analyse eine große Herausforderung dar, da hier viele Parameter manuell eingestellt werden können, die sich auf das Ergebnis auswirken. Die gewählten Einstellungen, sowie die Extraktion von Rekombinationen beziehungsweise unbestimmten Basen, werden oft nicht im Detail in Publikationen beschrieben, wodurch die Reanalyse der Daten erschwert wird[162]. Diese Probleme treffen auch auf Anwendungen zu, die selbst erstellte Assemblierungen der Nutzer für ihre Analysen verwenden. Sobald Analysen manuelle Schritte beinhalten oder wichtige Einstellungen vorgenommen werden müssen, fordert dies Nutzer ohne bioinformatisch Kenntnisse heraus. Falsch eingestellte Parameter oder die Wahl einer genomisch zu weit entfernten Referenzsequenz wirken sich deutlich auf die resultierenden genomischen Distanzen aus. Die cgMLST-Analyse in EnteroBase bietet demnach eine reproduzierbare und leicht durchführbare Alternative zu der SNP-Analyse.

4.1.3 Aussicht und Entwicklungspotential der Methoden

Die bioinformatischen Werkzeuge in EnteroBase könnten um die Möglichkeit erweitert werden eine Distanzmatrix der paarweisen Vergleiche zu berechnen, sodass die cgMLST-Analyse in Zukunft komplett dort durchführbar ist. Zwar können momentan die cgMLST Allelprofile in EnteroBase über eine API mit BioNumerics synchronisiert werden um dort eine Distanzmatrix zu generieren, jedoch ist so eine kommerzielle Software notwendig. Weiterhin könnten mit der Distanzmatrix Einträge zukünftig auch in ihrem akzessorischen Gengehalt verglichen werden. Zu diesem Zweck müsste neben dem cgMLST und dem

wgMLST Schema noch ein um die Kerngenom-Loci reduziertes wgMLST Schema verfügbar sein, welches nur die akzessorischen Bereiche beinhaltet.
Des Weiteren wäre es interessant *k-mer* basierte Methoden wie sie die Software PopPUNK[141] verwendet oder die hash-cgMLST[64] Methode in EnteroBase zu implementieren. Durch die Verknüpfung mit der EnteroBase Plattform wäre die Analyse von standardisierten, hoch qualitativen Assemblierungen durch diese Methoden gesichert und Nutzer könnten diese über eine graphische Oberfläche bedienen. Der Ansatz der hash-cgMLST Methode, jede individuelle Allelsequenz in einen individuellen *Hash* zu überführen könnte eine Alternative zu den momentan verwendeten Alleldatenbanken in EnteroBase sein, da diese zur Qualitätssicherung intensiv kuratiert werden müssen. Zudem benötigen *Hashes* weniger Speicherplatz als Allelsequenzen. PopPUNK wiederum würde die Berechnung des akzessorischen Gengehalts direkt in EnteroBase ermöglichen und durch flexibles Bestimmen der akzessorischen Genomabschnitte die Schema-abhängige Analyse ergänzen.
Da aber auch die cgMLST-Analyse Nachteile gegenüber der SNP-Analyse zeigte, wäre es von großem Interesse die Durchführung der SNP-Analyse zu standardisieren und zu vereinfachen. Neben der assemblierungsbasierten SNP-Analyse in EnteroBase ist die ASA^3P Pipeline[78] eine weitere Option für SNP-Analysen von nahen genomischen Verwandtschaften zwischen Isolaten. Die Pipeline assembliert die Illumina-*Reads* und vergleicht diese anschließend miteinander. Allerdings handelt es sich auch hier um eine assemblierungsbasierte SNP-Analyse, die, wie in Kapitel 4.1.2 diskutiert, durch mögliche Misassemblierungen Nachteile gegenüber der *Read-Mapping* basierten SNP-Analyse zeigt. Die Entwicklung einer Pipeline ähnlich der ASA^3P Pipeline mit bioinformatischen Werkzeugen für das *Read-Mapping* gegen eine Referenzsequenz und des anschließenden *Variant-callings* könnte eine standardisierte SNP-Analyse ermöglichen. Die Implementierung solcher bioinformatischer Werkzeuge in EnteroBase ergibt zum jetzigen Zeitpunkt keinen Sinn, da die Datenbank nur die erstellten Assemblierungen beinhaltet und die zugehörigen Illumina-*Reads* aus Speicherplatzgründen nicht hinterlegt werden. Allerdings gibt es erste Ideen die *Reads* anhand eines *Hash*-Algorithmus in *Hashes* zu überführen um so die Dateigröße der *Read*-Dateien zu minimieren. Vorrangig hat diese Überlegung zum Ziel, die den Assemblierungen zugrunde liegenden *Reads* zu archivieren und dadurch Duplikate in der Datenbank erkennen zu können. Die *Reads* würden so allerdings nicht mehr die genaue Basenabfolgen beinhalten und eine klassische *Read-Mapping* basierte SNP-Analyse wäre anhand der *Hashes* nicht möglich. In wieweit in *Hashes* umgewandelte Illumina-*Reads* für mappingbasierte Analysen genutzt werden können müsste noch näher untersucht werden. Eine Detektion von Punktmutationen anhand von *Hashes* ist allerdings auszuschließen.

4.2 Kurzzeitige Evolution und örtliche Ausbreitung von *C. difficile*

Die *C. difficile* Datenbank in EnteroBase ermöglicht erstmals einen umfangreichen Überblick über die Populationsstruktur des Pathogens und dessen globale Verteilung. Die einheitliche Analyse der Sequenzdaten, die cgMLST Typisierung und die anschließende hierarchische Clusterung der assemblierten Genome vereinfachen die Kommunikation zwischen Wissenschaftlern sowohl innerhalb eines Landes als auch länderübergreifend. Wie EnteroBase zur effektiven Detektion von pandemischen und endemischen Stämmen, wie aber auch zur Ausbruchsdetektion verwendet werden kann, wird in den folgenden Kapiteln diskutiert.

4.2.1 Erfassung der globalen Populationsstruktur auf verschiedenen Ebenen

Pandemische Stämme fallen in ein HC10 Cluster

Durch die hierarchische Clusterung auf HC10 Ebene konnten publizierte Pandemien mittels der cgMLST-Analyse nachgebildet werden. Hinweise auf sich ausbildende, pandemische *C. difficile* Stämme könnten demnach in Zukunft frühzeitig durch die Clusterung der Genome in HC10 Cluster erkannt

werden. Durch die Datenbank in EnteroBase werden automatisch alle Einträge in die Analyse mit einbezogen, wodurch Pandemien nicht mehr nur anhand eines begrenzten Datensatzes untersucht werden müssen. Die in der Literatur beschriebenen pandemischen Stämme von *C. difficile* wurden aufgrund ihrer phylogenetischen Clusterung in Untergruppen eingeteilt, die für die RT017 Pandemie anhand der HC10 Cluster nicht nachgebildet werden konnte (Abbildung 3.2, Seite 31). Während die fluorchinolonresistenten RT027 Untergruppen durch Äste getrennt wurden, die einer SNP Distanz von 7-10 entsprechen, entsprach die Astlänge zwischen den beiden Untergruppen der RT017 Pandemie einer SNP Distanz von 4[15], [103]. Die geringe SNP Distanz ließ den Schluss auf eine entsprechend geringe Anzahl an Kerngenom-Allelunterschieden zwischen den Genomen der beiden RT017 Untergruppen zu, wodurch sich diese anhand der HC10 Clusterung nicht in zwei unterschiedliche Gruppen aufteilen ließen. Die beiden Untergruppen werden in der Publikation mit unterschiedlichen Varianten der Toxingene in Verbindung gebracht, zeigen aber sonst keine weiteren charakteristischen phänotypischen Eigenschaften. Dem gegenüber spiegeln die Untergruppen FQR1 und FQR2 der RT027 Pandemie jeweils die Aufnahme der Fluorchinolonresistenz in zwei sich separat entwickelnden *Lineages* wider[15]. Biologisch gesehen bedeutet dies, dass sich Resistenzen gegen Antibiotika unabhängig voneinander in einem Organismus ausbilden und in verschiedenen Variationen auftreten können. Die Variationen der Toxingene bilden in den beiden Untergruppen keine biologisch bedeutenden Charakteristiken ab. Demnach trägt die Unterteilung der RT017 Pandemie nicht zu wichtigen biologischen Erkenntnissen bei.

Endemische Stämme können durch HC150 Cluster erfasst werden

In dieser Arbeit wurde erstmals die etablierte Typisierungsmethode der PCR Ribotypisierung mit der hierarchischen Clusterung von 13.515 *C. difficile* Genomen auf cgMLST Ebene verglichen. Die hierarchische Clusterung in HC150 Cluster stimmte größtenteils mit der PCR Ribotypisierung überein und deckte sich zudem besser mit den phylogenetischen Gruppen (Abbildung 3.4, Seite 33). Wurden multiple PCR Ribotypen in eine phylogenetische Gruppe zusammengefasst kann dies darauf zurückzuführen sein, dass die PCR Ribotypisierung anfälliger für Rekombinationen ist als die MLST-Analyse [48]. Eine Rekombination in den ITS-Regionen würde die Länge des DNA-Fragments ändern und dadurch zu einem anderen Bandenmuster führen. Das Isolat würde also einem anderen PCR Ribotypen zugeordnet werden. Auch in der cgMLST-Analyse wirken sich Rekombinationen, wenn sie denn überhaupt im Kerngenom auftreten, auf die Länge der Allelsequenzen aus. Dadurch wird die Allelsequenz einer anderen Allelnummer zugeordnet und sich somit zu einem Kerngenom-Allelunterschied führen. Bei einer erlaubten kettenweisen genomischen Distanz von 150 Kerngenom-Allelunterschieden in einem HC150 Cluster würde das Isolat durch einen zusätzlichen Unterschied wahrscheinlich keinem anderen HC150 Cluster zugeordnet werden und die Typisierung nicht beeinflussen.
Ein weiterer Grund ist die Verwendung von abweichenden Nomenklaturen für die PCR Ribotypisierung in verschiedenen Laboratorien. Dies kann dazu führen, dass identische oder sich stark ähnelnde Bandenmuster anderen PCR Ribotypen zugeordnet werden. Die zugehörigen Genome würden in eine phylogenetische Gruppe fallen, durch die PCR Ribotypisierung aber unterschiedlich typisiert worden sein [59].
Die unterschiedlichen Nomenklaturen könnten natürlich auch das genaue Gegenteil bewirken und die Verteilung von gleich bezeichneten Genomen über mehrere phylogenetische Gruppen erklären. Zudem betraf dies PCR Ribotypen, die sich nicht eindeutig bestimmen lassen und oft in Untergruppen klassifiziert werden. So wurden zum Beispiel Isolate des PCR Ribotyps 014 durch die EC PCR Ribotypisierung in sieben Untergruppen aufgeteilt [54]. Diese sieben Untergruppen fielen nicht in sieben voneinander getrennte phylogenetische Gruppen (hier nicht gezeigt) und deuten eine, wie von Indra et al. schon vermutet[54], Überdiskriminierung der Isolate durch die EC PCR Ribotypisierung an.
Das die PCR Ribotypisierung von den phylogenetischen Gruppen abweicht könnte weiterhin daran liegen, dass das cgMLST Schema nicht zwingend alle ITS-Regionen, die in der PCR Ribotypisierung amplifiziert werden, umfasst[154]. Nahe beziehungsweise distinkte Verwandtschaften, die aus der PCR Ribotypisierung gefolgert werden, könnten auf genomische Abschnitte begründet sein, die in der cgMLST-Analyse nicht

berücksichtigt werden. Hier sei allerdings angemerkt, dass die cgMLST-Analyse Genome in weitaus mehreren konservierten Genomabschnitten miteinander vergleicht und somit zu einer höheren Auflösung führt.
Da viele der 201 HC150 Cluster kein Genom mit PCR Ribotypeninformation umfassten, konnte der PCR Ribotyp nur für einen gewissen Teil der 13.515 Genome vorhergesagt werden. Die Bezeichnung der distinkten phylogenetischen Gruppen als cgST Complexe (CCs) würde dieses Problem umgehen und eine einheitliche Nomenklatur für die auf den HC150 Clustern basierende Typisierungsmethode bieten. Ein neuer Eintrag in EnteroBase wird automatisch den bestehenden CCs oder einer neuen CC-Nummer zugeordnet. Neu aufkommende endemische Stämme können dadurch direkt erkannt und einheitlich bezeichnet werden. Bei der PCR Ribotypisierung stellt ein von den Referenzmustern abweichendes Bandenmuster ein Problem dar und kann meist nicht zugeordnet werden[48].

Die Diversität der Populationsstruktur

Trotz der umfangreichen Anzahl an *C. difficile* Genomen in der Datenbank, wurden nur zwei Drittel der genomischen Diversität durch die untersuchten Einträge erfasst (Abbildung 3.6, Seite 36). Die große Anzahl an Genomen, die sich keinem HC150 Cluster zuordnen ließen, unterstützt diese Aussage (209 Singletons, Kapitel 3.2.2, Seite 31). Diese Genome könnten die Grundlage für zukünftig neu auftretende endemischen Stämme bilden. Viele dieser Singletons fallen in dem in Abbildung 3.5 (Seite 34) dargestellten phylogenetischen Baum auf denselben Ast wie annotierte HC150 Cluster und könnten somit endemischen Stämme repräsentieren, die sich von den schon bekannten CCs abspalten. Als Beispiel sei hier CC4 genannt, der mit multiplen Singletons und kleineren, teilweise nur aus zwei Genomen bestehenden HC150 Clustern ein Monophylum bildet. Hier haben sich demnach schon zum Teil neue HC150 Cluster formiert.
Isolate, die zu den bisher identifizierten kryptischen Kladen C-I, C-II und C-III gehören, stammen meistens aus der Umwelt und zeigen durch ihre große genomische Distanz zu anderen endemischen Stämmen von *C. difficile* keine direkte Assoziation zu menschlichen oder tierischen Isolaten[29]. Die hohe, zum Teil auch hier noch nicht komplett erfasste, Diversität von *C. difficile* könnte somit auf dessen breites Spektrum an Nischen zurückzuführen sein. Besonders die aus der Umwelt stammenden Isolate, deren Genome sich nicht einem HC150 Cluster zuordnen ließen, könnten native Stämme repräsentieren, die sich noch nicht in der tierischen oder menschlichen Population ausgebreitet haben[29]. Dass es neben den definierten kryptischen Kladen noch weitere distinkte *C. difficile* Stämme gibt wurde schon vermutet und die bisherige Unkenntnis über diese mit fehlenden Daten aus bestimmten geographischen Regionen begründet[27]. Die Analyse aller momentan global verfügbaren Sequenzdaten von *C. difficile* konnte zwar zusätzliche Kladen detektieren, scheint aber immer noch kein vollständiges Bild der Diversität des Bakteriums abzugeben.
Die bekannten Merkmale endemischer *C. difficile* Stämme wie PCR Ribotyp 027 (CC4), 078 (CC1) und 017 (CC17) spiegelten sich weitestgehend auch in den Analysen der Datenbank und den genomischen Beziehungen auf Kerngenom-Ebene wider. So wurde ein großer Anteil der Genome der HC150 Cluster CC4 und CC17, die in den globalen HC2 Clustern vorkamen, in einen großen HC2 Cluster zusammengefasst (Tabelle 3.4, Seite 44). Dies ließ für beide CCs auf ein Vorkommen in Form eines großen Ausbruchs schließen, der über viele Jahre bestehen blieb und sich über mehrere Länder verbreitete. Für die entsprechenden PCR Ribotypen wurde dieses Verhalten schon in früheren Studien vermutet, war aber allerdings immer auf einem begrenzten Datensatz begründet[15], [103]. Isolate des CC17 scheinen ihren Ursprung in Asien zu haben und sich von dort aus auszubreiten, da der entsprechende Ribotyp stetig in asiatischen Ländern registriert wurde[27]. Der Ursprung von CC4 Isolaten lässt sich laut früheren Studien in Nordamerika vermuten, wobei sich besonders der fluorchinolonresistente Stamm global verbreitete[15], [16]. Hierbei scheint die Einbringung beziehungsweise die folgende Ausbreitung des Stammes in einem Land von der Verabreichung entsprechender Antibiotika abzuhängen[80].
Ein anderes epidemiologisches Verhalten zeigten Isolate des HC150 Clusters CC1. Für diesen CC ließen sich die meisten globalen HC2 Cluster verzeichnen. Zudem bildeten die Genome, die in globalen HC2 Clustern vorkamen, kein dominantes HC2 Cluster aus (Tabelle 3.4, 44). Dies ließ auf wiederholte Ausbrüche durch

Isolate des CC1 schließen, die sich über mehrere Länder erstreckten. Der, im Vergleich zu den anderen CCs, hohe Anteil an tierischen Isolaten in CC1 lässt zudem eine Ausbreitung des Stammes über Nutztiere beziehungsweise Lebensmittel, insbesondere Fleisch, vermuten. Dieses Verhalten wurde schon in früheren Studien gezeigt, wobei hier besonders die genomische Ähnlichkeit von Isolaten aus Tieren und Menschen und deren Ausbreitung zwischen Nordamerika und Europa hervorgehoben wurde[87]. Wurden Isolate des entsprechenden PCR Ribotypen 078 früher hauptsächlich in Nutztieren nachgewiesen, so trat der Stamm immer häufiger in Menschen auf und entwickelte sich zum dritthäufigsten Stamm in Europa[81], [82]. Besonders die fehlende geographische Clusterung nah verwandter Genome und der Nachweis des Stammes auf Fleisch ließen eine Verbreitung über Lebensmittelvergiftungen vermuten. Ein endgültiger Beweis steht jedoch noch aus[81], [87]. Eine alternative Erklärung wären kontaminierte Produktionsherde und deren Umgebung[83], [84].

Liegt der Fokus in der Literatur oft auf der Untersuchung spezifischer PCR Ribotypen, so zeigte die Analyse der umfangreichen Datenbank in EnteroBase interessante Verhalten von *C. difficile* Stämmen auf, die bislang nicht groß beachtet wurden und die als Grundlage für weitere detailliertere Studien dienen könnten. So zeigten neben dem oft in tierischen Proben nachgewiesene CC1 auch CC22 (PCR RT106/500), CC88 (PCR RT014) und CC79 (RT010) einen kleinen Anteil an Isolaten aus tierischen Quellen (Tabelle 3.2, Seite 35). Während das Vorkommen des PCR Ribotypen 014 in Schweinen und Menschen schon näher untersucht wurde[73], liegen zu den PCR Ribotypen 106/500 und 010 noch keine detaillierteren Studien vor. Mit einem Bestehen von bis zu 20 Jahren und einem Anteil an Genomen, die globale HC2 Cluster bildeten, könnten diese Stämme eine mögliche Ursache für die globale Ausbreitung von *C. difficile* über tierische Produkte sein. Eine auf mehrere PCR Ribotypen ausgebreitete Analyse des epidemiologischen Verhaltens wurde 2018 von Eyre et al. veröffentlicht [167]. Für jeden Ribotypen wurde die länderbasierte Sequenzclusterung anhand der Medianwerte der paarweisen SNPs zwischen Isolaten innerhalb eines Landes und länderübergreifend unter Verwendung von Permutationstests bewertet. Hierbei wurden die Isolate systematisch für diese Studie gesammelt, um so aus jedem Land und für jeden Ribotypen ein repräsentatives Set zu haben. Auch wenn die getätigten Aussagen in dieser Studie größtenteils mit den Ergebnissen dieser Arbeit übereinstimmen, konnte die Analyse durch die cgMLST-Analyse der Datenbank in EnteroBase nicht nachgebildet werden. Dadurch, dass alle veröffentlichten Sequenzdaten in die Datenbank mit eingebracht werden, ist die Anzahl an Isolaten eines Ribotypen und deren geographischer Ursprung stark von den durchgeführten und publizierten Studien zu *C. difficile* abhängig. Die Auswahl eines repräsentativen Satzes an Isolaten aus jedem Land für jeden Ribotypen hätte zufällig erfolgen müssen und wäre somit nicht vergleichbar mit Eyre et al. Studiendesign gewesen.

Mit der Clusterung der Genome nach HC2000 konnten vier der bekannten fünf Kladen von *C. difficile* nachgebildet werden (Anhang B; Abbildung B.2). Allerdings fasste die auf paarweisen Kerngenom-Allelunterschieden basierende HC200 Clusterung die von Dingle et al.[152] beschriebenen Kladen 1 und 2 in ein HC2000 Cluster zusammen, obwohl die Phylogenie des Baumes das Abspalten der zweiten Klade vermuten lässt (Abbildung 3.5, Seite 34). Hierbei wird erneut der schon in Kapitel 4.1.1 beschriebene Unterschied zwischen der phylogenetischen Distanz und der auf der Anzahl an Punktmutationen beziehungsweise Allelunterschieden basierenden Distanz und der daraus resultierenden Clusterung der Genome deutlich. Die auf der Phylogenie basierenden genomischen Distanzen könnten, besonders bei hoch diversen Datensätzen, größere Unterschiede zwischen Isolaten andeuten, als die tatsächliche Anzahl der Punktmutationen beziehungsweise Allelunterschiede wiedergibt. Die Genome würden demnach in eine deutlich abgegrenzte phylogenetische Gruppe fallen, wobei die hierarchische Clusterung die Genome zusammen in ein Cluster fasst. Die Phylogenie sollte besonders dann berücksichtigt werden, wenn die Bildung einzelner Kladen und deren gemeinsame Vorfahren untersucht werden sollen. Durch das Verständnis der lang zurückreichenden Evolution eines Organismus kann dessen mögliches zukünftiges Verhalten eingeschätzt werden. Wenn allerdings ein direkter paarweiser Vergleich von Genomen vorgenommen werden soll, ist die Anzahl an Mutationen vorzuziehen.

Die hierarchische Clusterung scheint auch auf dieser Populationsebene mit dem HC2000 Cluster eine Alternative zu den klassisch verwendeten 5 Kladen zu bieten. Die Bildung einer weiteren Klade zwischen den Kladen 3 und 4 unterstützt diese Aussage, da unter Einbeziehen einer Vielzahl an Genomen die Einteilung zumindest im Detail nicht mehr aktuell scheint. Die Vermutung von Knetsch et al. [48], dass eine höher auflösende Typisierungsmethode die hoch diverse Klade 1 weiter unterteilen könnte, wurde mit der Anwendung der hoch auflösenden cgMLST-Analyse zumindest auf HC2000 Ebene widerlegt.

Die hier getätigten Aussagen basieren auf den in EnteroBase verfügbaren Metadaten, die teilweise manuell hinzugefügt, teilweise über NCBI bezogen wurden. Hierbei ist die Interpretation der Daten stark von der Korrektheit der angegebenen Daten abhängig. Bei Einträgen, die einer Publikation angehören, können die Metadaten durch intensive Kuratierung überprüft werden. Dies ist allerdings bei von Nutzern hochgeladenen Daten, sowie bei Daten aus NCBI, die keiner Publikation angehören, nicht möglich. Hier muss auf das gemeinsame Interesse der Nutzer an einer vollständigen und korrekten Datenbank von *C. difficile* vertraut werden. Des Weiteren sind in der Datenbank Länder, in denen viele Studien zu *C. difficile* durchgeführt werden, stark überrepräsentiert (Vereinigtes Königreich, Australien, Nordamerika, Deutschland) und andere Länder von Kontinenten, auf denen sich Untersuchungen eher auf einzelne Krankenhausstudien begrenzen (zum Beispiel Südamerika, Afrika) stark unterrepräsentiert beziehungsweise nicht vorhanden. Dadurch wird ein geographischer Überblick der *C. difficile* Population verzerrt dargestellt und Ausbreitungen zwischen Ländern können nur begrenzt erfasst werden. Trotzdem bietet die Datenbank mit ihrer umfangreichen Ansammlung an *C. difficile* Genomen und den dazugehörigen Metadaten eine Möglichkeit, die Populationsstruktur des Bakteriums in einem neuen Umfang zu analysieren und Tendenzen von neu aufkommenden Stämmen im globalen Kontext zu erkennen.

4.2.2 Detektion von nah verwandten Isolaten

Rezidiv versus Neuinfektion

Durch die cgMLST-Analyse konnten rezidivierende CDI eindeutig von Neuinfektionen durch einen neu aufgenommenen Stamm unterschieden werden. Genome von Isolaten, die aus zwei Episoden eines Patienten mit rezidivierender CDI isoliert wurden, fielen einheitlich in ein HC2 Cluster (Abbildung 3.9 Patienten G, D und F; Seite 39). Dies entspricht den bisherigen Annahmen einer nahen genomischen Verwandtschaft von $\leq$2 SNPs für Isolate eines Rezidivs[40]. Ebenso zeigten die Isolate, die aus einem Patienten mit einer Neuinfektion isoliert wurden, eine deutliche genomische Distanz zueinander, die den angegebenen Wert einer SNP Distanz von $>$ 10 für eine Neuinfektion deutlich überstieg[40] (Abbildung 3.9 Patient E). Bisherige Studien beruhen meist auf jeweils einem Isolat aus den Krankheitsepisoden der Patienten und berichten von einem hohen Anteil (76 %[40]) an rezidivierender CDI, bei denen das infektiöse Bakterium nach erfolgreicher Behandlung den Darm des Patienten weiterhin besiedelte und dadurch eine erneute Infektion auslösen konnte. Allerdings könnten wiederkehrenden Infektionen auch durch Aufnahme eines identischen Stammes aus der kontaminierten Umgebung des CDI Patienten verursacht werden und so den Fall eines Rezidivs vortäuschen[38].
Die initiale Infektion eines Rezidivs scheint auch durch mehrere, nah verwandte Stämme auslösbar zu sein (Abbildung 3.9 Patient F). Nimmt man den in der Literatur vorgeschlagenen Grenzwert von $\leq$2 genomischen Unterschieden für eine rezidivierende Infektion an, so war dieser Patient mit fünf verschiedenen *C. difficile* Stämmen besiedelt. Da die Isolate der zweiten Episode genomisch allerdings nur einem dieser Stämme ähneln, konnten die anderen vier vermutlich durch die Behandlung während der ersten Episode verdrängt werden. Dieses Beispiel zeigt, dass die Analyse von jeweils nur einem Isolat aus beiden Episoden nicht ausreicht um Fälle in Gänze zu erfassen. Zwei der Isolate aus der ersten Episode zeigen einen genomischen Unterschied von $\geq$10 zu den anderen Isolaten. Anhand dieser Isolate wäre der Patient mit einer Neuinfektion diagnostiziert worden, obwohl andere Isolate der ersten Episode eine nahe genomische Verwandtschaft zu den Isolaten aus der zweiten Episode aufzeigen und demnach auf ein Rezidiv schließen lassen.

Da die Unterscheidung zwischen Rezidiv und Neuinfektion im klinischen Alltag nur schwer möglich ist, werden die Fälle von Patienten mit wiederkehrender CDI anhand des Zeitpunktes beurteilt. Dabei wird der Patient mit einem Rezidiv diagnostiziert, wenn die beiden Episoden < 8 Wochen auseinander lagen. Eine Infektion durch einen neu aufgenommenen Stamm wird demnentsprechend bei einer Zeitspanne zwischen den beiden Episoden von > 8 Wochen festgestellt[168], [169]. Wie auch schon in anderen Studien berichtet, übersteigt die zeitliche Differenz der beiden Episoden der in dieser Arbeit durch genomische Analysen diagnostizierte rezidivierenden CDI mit 22 Wochen die 8 Wochen deutlich[41], [155]. Demnach liegt der Verdacht nahe, dass anhand des zeitlich festgelegten Grenzwertes in der Vergangenheit eine Vielzahl an rezidivierenden CDI als Neuinfektion diagnostiziert wurden. Somit sollte eine Erweiterung dieses Grenzwertes in Erwägung gezogen werden.
Die Differenzierung von wiederkehrenden CDI in Neuinfektionen oder Rezidive ist von hoher Bedeutung, da hieraus wichtige Erkenntnisse für Maßnahmen für das Management von Patientenverlegungen und der Krankenhaushygiene gewonnen werden können[155]. Besonders herausfordernd wird es, wenn der betroffene Patient in einem anderen Krankenhaus mit erneuter CDI diagnostiziert wurde. Neben zwei unerkannten rezidivierenden Fällen im gleichen Krankenhaus konnte in dieser Arbeit durch retrospektive Epidemiologie ein Patient mit rezidivierender CDI aufgezeigt werden, dessen zweite Episode in einem anderen Krankenhaus als neu auftretende Infektion diagnostiziert wurde (Kapitel 3.4).

Ausbruchsanalysen

Neben der Detektion von rezidivierenden Fällen konnten durch die cgMLST-Analyse der Genome von Isolaten, die aus einem Netzwerk von Krankenhäusern isoliert wurden, retrospektiv mögliche Transmissionswege innerhalb wie auch zwischen den beprobten Krankenhäusern aufgedeckt werden. Das Herstellen eines Zusammenhangs zwischen insgesamt 66 Patienten demonstriert, dass die „umgekehrte genomische Epidemiologie“ zur Ausbruchsdetektion angewendet werden kann. Dabei wird bei der umgekehrten genomischen Epidemiologie bei einer ausreichend nahen genomischen Verwandtschaft von zwei Isolaten angenommen, dass diese eine Folge einer Infektion aus einer gemeinsamen Quelle sind[165]. Als nahe genomische Verwandtschaft wurde hier ein Wert von ≤ 2 Kerngenom-Allelunterschieden angenommen, entsprechend dem von Eyre et al. bestimmten Grenzwert von ≤ 2 SNPs für eine mögliche Transmission von *C. difficile* Isolaten, da die cgMLST-Analyse wie in Kapitel 4.1.2 diskutiert vergleichbare genomische Distanzen wie die SNP-Analyse lieferte.
Insgesamt fielen 133 der 290 in die Analyse mit eingegangenen Isolate in HC2 Cluster und besaßen somit jeweils eine kettenweise genomische Distanz von höchstens 2 Kerngenom-Allelunterschieden zueinander. Allerdings konnte nicht für alle ein epidemiologischer Zusammenhang, wie beispielsweise ein Aufenthalt auf der gleichen Station, nachgewiesen werden. Neben fehlender oder unvollständiger Dokumentation der Patientendaten ist dies auch auf das Studiendesign zurückzuführen (Kapitel 2.1.1, Seite 14). Neben der Berücksichtigung der lediglich ersten 20 Fälle pro Periode, stieg Krankenhaus 6 nach zwei Perioden aus der Studie aus und wurde durch Krankenhaus 4 ersetzt. Dadurch konnten mögliche Transmissionswege, die Patienten in Krankenhaus 6 betrafen, in der dritten Periode nicht mehr erfasst werden. Dies betrifft sowohl das Verlegen eines an CDI erkrankten Patienten aus Krankenhaus 6 in eines der anderen beprobten Krankenhäuser als mögliche Quelle einer Transmission, als auch mögliche Akzeptoren einer CDI in Krankenhaus 6 durch einen erkrankten Patienten aus den anderen Krankenhäusern. Des Weiteren lag oft nur das Krankenhaus und nicht die Information zur genauen Station vor, sodass mögliche Zusammenhänge nur auf Krankenhausebene erfasst werden konnten. Detailliertere Stationsinformationen hätten unter Umständen zu einer ähnlich hohen Signifikanz der Assoziation zwischen Station und Bildung der HC2 Cluster geführt, wie sie für die Krankenhäuser erzielt wurde (Kapitel 3.4.2, Seite 40). Weiterhin zeigte sich bei der Analyse der epidemiologischen Daten wie wichtig die Dokumentation des kompletten Verlaufs des Patienten ist. Je umfangreicher die Erfassung der Krankenhausaufenthalte vor und nach der Diagnose der CDI waren, desto eher konnte ein Zusammenhang zu einem anderen Patienten hergestellt werden. Dies betraf vor

allem mögliche indirekte Transmissionswege, da hier die Stationen auf denen die Patienten diagnostiziert wurden meist nicht miteinander übereinstimmen. Die Dokumentation des genauen Diagnosedatums ist weiterhin eine essenzielle Information um wiederum direkte Transmissionswege nachbilden zu können. Eine nur auf genomischen Distanzen beziehungsweise phylogenetischen Untersuchungen basierte Analyse von Transmissionen kann demnach nur begrenzt stattfinden. Die Kombination mit epidemiologischen Daten ist hier unumgänglich[107].

Wie in den Ergebnissen in Kapitel 3.4.2 schon andiskutiert, können neben fehlenden epidemiologischen Daten und begrenztem Umfang der Studie auch asymptomatische Patienten als potentielle Übertragungsquelle in Betracht gezogen werden. Besonders bei der Studie von García-Fernández et al. ist dies eine legitime Erklärung für das Aufteilen der Genome in multiple HC2 Cluster, da das Studiendesign die Erfassung aller an CDI erkrankten Patienten in der Beprobungszeit vorsah[55].
In der Literatur findet man immer wieder Studien, die der Erfassung von asymptomatischen Patienten eine wichtige Rolle zuschreiben. So konnte durch ein generelles Screening bei der Patientenaufnahme und der anschließenden Isolierung bei positivem Befund auf CDI trotz fehlender Symptome die Anzahl an nosokomialen CDI-Fällen in einem Krankenhaus signifikant gesenkt werden[170]. So konnte nachweislich die Einbringung von *C. difficile* Isolaten aus der Gemeinschaft in das Krankenhaus eingedämmt werden. In einer anderen Studie wurde eine Verdopplung der CDI-Fälle durch Hospitalisation von Patienten mit einem asymptomatischen CDI Patienten auf dem gleichen Zimmer festgestellt[25]. Dem gegenüber steht eine Studie von Eyre et al., die eine Transmission durch asymptomatische Patienten als eher unwahrscheinlich ansieht[171]. Allerdings wird auch hier auf die hohe Anzahl an CDI Patienten ohne Symptome und die dadurch entstehende potentielle Quelle für weitere Transmissionen hingewiesen. Aktuell rät sowohl das Robert Koch-Institut als auch das Europäische Zentrum für die Prävention und die Kontrolle von Krankheiten (ECDC) von einem Aufnahmescreening auf CDI ab[14], [172]. Jedoch sollte ein CDI-Test von Zimmernachbarn einer infizierten Person in Betracht gezogen werden, um eine mögliche Ausbreitung des Pathogens durch Patientenverlegung zu verhindern.
Ergänzend zu den vorgebrachten Argumenten sei hier auch auf die Möglichkeit von methodischen Ursachen hingewiesen. In der Korrelationsanalyse der Kerngenom-Allelunterschiede und SNP Distanzen in Kapitel 3.3 zeigten die Datensätze der vier publizierten Ausbrüche die geringste Korrelation zwischen den Distanzwerten (R^2=0,71; Abbildung 3.7, Seite 37). Wie schon in Kapitel 4.1.2 diskutiert, werden in der cgMLST-Analyse auch Insertionen und Deletionen berücksichtigt, sodass im Vergleich zu den SNP Distanzen höhere Kerngenom-Allelunterschiede vorkommen können. Daran könnte erklärt werden, warum sich die Ausbruchsisolate über multiple HC2 Cluster erstrecken (Abbildung 3.10, Seite 40). Weiterhin wurden in die Reanalyse der vier Ausbrüche nur die in den Publikationen als ein Ausbruch definierten Isolate mit einbezogen. Da es aber auch in der SNP-Analyse zur Detektion von falsch-positiven Mutationen kommen kann, könnten die in der Transmissionskette fehlenden Isolate durch fälschlicherweise zu hoch berechnete SNP Distanzen in den Publikationen vom Ausbruch ausgeschlossen worden sein. Eine cgMLST-Analyse aller Genome der in den Studien isolierten Isolate hätte unter Umständen zur Füllung der Lücken in der Transmissionskette und somit zur einheitlichen Clusterung der Genome in ein HC2 Cluster geführt.

Als Teil des SOARiAL Projektes erbrachte die vorliegende Arbeit erstmals den Beweis der Kontamination von landwirtschaftlich genutzten Flächen durch ausgebrachten und eingearbeiteten Dung aus Geflügelmist von *C. difficile* Sporen und deren Überleben in dem gedüngten Boden über mehrere Wochen. Dabei wurden in dem gedüngten Boden ausschließlich *C. difficile* Stämme nachgewiesen, die den gleichen HC150 Clustern wie die Isolate aus dem ausgebrachten Mist angehören. Eine zusätzliche Analyse der HC2 Cluster bestätigte die nahe genomische Verwandtschaft zwischen diesen Isolaten und deutete auf eine Transmissionskette hin, die ihren Ursprung in dem ausgebrachten Geflügelmist hat und über 19 Wochen in dem gedüngten Boden bestehen blieb. Obwohl nur ein einzelnes Staub-Isolat aus dem Staub, der während des Windkanalversuchs von dem gedüngten Boden aufwirbelte, isoliert wurde, zeigt dessen nahe genomische Verwandtschaft zu den

Bodenisolaten die prinzipielle Möglichkeit einer Austragung von *C. difficile* von dem gedüngten Boden auf den Staub.

Das der ausgebrachte und eingearbeitete Dung Einfluss auf die mikrobielle Gemeinschaft in dem Boden hat wurde schon anhand von 16S rRNA Sequenzdaten für das Vorkommen von verschiedenen Genera an Bakterien gezeigt[117]. Auch hier konnten Bakterien, die dominant in den Mistproben vorkamen, noch nach langer Zeit in dem gedüngten Boden und in Staubproben nachgewiesen werden, obwohl sie im ungedüngten Boden nur minimal vorkamen. Die Ergebnisse dieser Arbeit untermauern demnach die schon in der Literatur oft vermutete und in Thiel et al. auf Genera-Ebene nachgewiesene Verbreitung von Bakterien aus dem Mist, über den Boden in den Staub.[95], [96], [117] Die Staubpartikel sind dabei, ähnlich wie Aerosole, als mögliche Übertragungsmedien der Bakterien von dem gedüngten Boden auf den Menschen anzusehen.

Auch wenn in den hier untersuchten Proben die meisten Isolate dem HC150 Cluster CC3 zugeordnet wurden, konnte aufgrund der schwankenden Anzahl an Isolaten pro Probe keine wirkliche Aussage über einen dominierenden Stamm getroffen werden. Auch frühere Studien konnten in Geflügelmist keinen dominanten *C. difficile* Stamm nachweisen[89]. Weiterhin wurde bei einem Vergleich von *C. difficile* Isolaten aus unterschiedlichen Quellen eine große Diversität an PCR Ribotypen in Umweltproben festgestellt, wobei sich hier jedoch 11 der 90 identifizierten Ribotypen in allen untersuchten Quellen nachweisen ließen. Dadurch wurde die Fähigkeit von *C. difficile*, sich jeglichen Umgebungen anpassen zu können, unterstrichen[173]. Zudem wird demnach deutlich, dass *C. difficile* sich ubiquitär ausbreiten kann und dabei viele verschiedene Nischen einnimmt. Die hohe Diversität an PCR Ribotypen in der Natur könnte mit dem vermehrten Vorkommen von kryptischen Kladen, beziehungsweise von Singletons in EnteroBase in Verbindung gebracht werden (siehe Abbildung 3.5, Seite 34). Wie in Kapitel 4.2.1 schon diskutiert, können diese als mögliche native Stämme angesehen werden, die potentielle neue endemische Stämme darstellen.

Zieht man in Erwägung, dass laut des Robert Koch-Instituts Isolate des Ribotyps 001 (hier CC3) bei nosokomialen Infektionen in Deutschland überwiegen, könnte man wiederum einen Zusammenhang zwischen den hier untersuchten Isolaten und der Gesellschaft in Deutschland sehen (Stand: 2.2.2018): Das dominante Vorkommen von CC3 Isolaten sowohl im Mist als auch über längere Zeit im gedüngten Boden und die nahe genomische Verwandtschaft zu dem Staubisolat könnten der Ursprung des am häufigsten nosokomial vorkommenden Stammes sein, indem die pathogenen *C. difficile* aus dem Mist, über den Boden auf Staubpartikeln in die Gesellschaft getragen wurden. Unterstrichen wird dies durch die nahe genomische Verwandtschaft zwischen Isolaten aus dem gedüngten Boden und klinischen Isolaten, die Teil der *C. difficile* Datenbank in EnteroBase sind und nicht im Zusammenhang mit dieser Studie isoliert wurden. Diese Vermutung stützt sich allerdings nur auf eine kleine Zahl an Isolaten und benötigt zusätzliche Untersuchungen.

Sowohl Isolate der regional begrenzten Krankenhausstudie als auch Isolate aus dem SOARIAL Projekt wurden in HC2 Cluster mit studienfremden Isolaten geclustert, die keinen epidemiologischen Zusammenhang zeigten und in verschiedenen Ländern isoliert wurden. Dieses überraschende Ergebnis motivierte zu einer Analyse der kompletten *C. difficile* Datenbank in EnteroBase auf genomische Verwandtschaften zwischen epidemiologisch nicht zusammenhängende Isolate und deckte, neben den methodischen Herausforderungen die in Kapitel 4.1 diskutiert wurden, Limitierungen der Ausbruchsdetektion auf.

4.2.3 Limitierungen der Ausbruchsdetektion

Bei der cgMLST-Analyse von nahen genomischen Verwandtschaften zwischen epidemiologisch nicht miteinander verbundenen Isolaten wurden zum Teil große Unterschiede im akzessorischem Gengehalt festgestellt, wodurch die Schlussfolgerung auf eine mögliche Transmission der Isolate implausibel erschien. Hierbei handelte es sich um Einträge in EnteroBase, deren Isolate nicht aus dem gleichen Land stammten und deren Genome sich in $\leq$2 Kerngenom-Allelen unterschieden.

Eine umfassende Analyse der *C. difficile* Population auf mögliche Transmissionsketten zwischen epidemiologisch nicht verbundenen Isolaten war aufgrund der Reduzierung der Genome auf ihre cgMLST

Allelprofile durchführbar. Eine SNP-Analyse mit 13.515 Genomen ist rechnerisch kaum zu bewältigen[68], [114]. Deshalb liegen bis jetzt auch nur Publikationen vor, die eine globale Verbreitung von *C. difficile* auf Basis eines begrenzten Datensatzes untersuchten[56], [87]. Dennoch wurden geographisch auseinander liegende Isolate miteinander in Verbindung gebracht.

Das die nahe genomische Verwandtschaft zwischen den in dieser Arbeit detektierten Isolaten kein durch die cgMLST-Analyse hervorgerufenes Artefakt ist, bestätigen die Ergebnisse der SNP-Analyse der 816 ausgewählten Genome (45,71 % $\leq$2 SNPs; Tabelle 3.9, Seite 54). Isolate aus unterschiedlichen Ländern, deren Genome sich möglicherweise in $\leq$2 SNPs unterscheiden, jedoch >2 Kerngenom-Allelunterschiede zeigten, wurden durch die initiale cgMLST-Anlayse, die zur Vorauswahl der 816 Isolate führte, nicht erfasst. Die Anzahl an epidemiologisch nicht verwandten Isolaten mit einer genomischen Distanz $\leq$2 SNPs könnte demnach auch höher als 373 sein, erfordert aber SNP-Analysen der kompletten Datenbank.

Eine im Vergleich zum Kerngenom höhere genomische Distanz im akzessorischem Gengehalt zwischen Isolaten wurde sowohl in den Studien, die internationale Isolate über eine Transmissionskette verknüpften[56], als auch für pandemische Stämme festgestellt[103], jedoch nicht mit in die epidemiologische Schlussfolgerung einbezogen. Die geringe Anzahl an Mutationen im Kerngenom zwischen Isolaten, deren Probennahme sich sowohl zeitlich als auch örtlich stark voneinander unterscheiden, wird mit der Fähigkeit von *C. difficile* begründet, Endosporen zu bilden[56]. Der von Eyre et al. definierte Grenzwert für paarweise genomische Distanzen, die mit einer hohen Wahrscheinlichkeit eine Verbindung der Isolate durch eine Transmissionskette schlussfolgern lassen, wurde anhand eines Modells bestimmt, das auf den genomischen Distanzen zwischen Isolaten aus einem rezidivierenden Patienten beruht[72]. Des Weiteren wurde für das Modell eine molekulare Uhr angelegt, für die eine konstante Rate der Evolution angenommen wurde, wodurch die durch die Sporenbildung auftretende ruhende Natur des *C. difficile* Genoms nicht berücksichtigt und die eigentliche evolutionäre Distanz zwischen den Isolaten unterschätzt werden kann[56]. Dadurch könnten durch das Anlegen des Grenzwertes Isolate einer gemeinsamen Transmissionskette zugeordnet werden, obwohl es keinen direkten Zusammenhang gibt. Die teilweise epidemiologisch widersprüchlichen Daten der vorausgewählten Genome, also die zeitlichen und örtlichen Unterschiede der Isolation des entsprechenden Isolats, unterstützen dies.

Das Auftreten von ruhenden Sporen erklärt zwar die geringe Mutationsrate im Kerngenom von epidemiologisch nicht zusammenhängenden Isolaten, allerdings nicht die starken Unterschiede in den akzessorischen Bereichen. Zum einen könnte man vermuten, dass die geringe Anzahl an Kerngenom-Allelunterschieden auf die stochastische Natur von Mutationen zurückzuführen ist. Demnach stellen die für Isolate mit unterschiedlichem Herkunftsland berechneten genomischen Distanzen von $\leq$2 Kerngenom-Allelunterschieden lediglich die äußersten Ränder der statistischen Verteilung der Mutationshäufigkeiten dar. Zum anderen liegt die in den Berechnungen von Eyre et al. verwendete Mutationsrate über der, die für Mutationen in *C. difficile* Genomen über eine lange Zeit berechnet wurde[140]. Es wird angenommen, dass *C. difficile* im Laufe der Zeit schädliche Mutationen beseitigt und sich dadurch im Kerngenom wieder zurück entwickelt[27]. Das akzessorische Genom wäre davon nicht betroffen. Demnach würde eine Untersuchung der Kerngenome eine nahe genomische Verwandtschaft andeuten, die allerdings durch die über eine lange Zeitspanne auftretende Bereinigung der Mutationen im Kerngenom verursacht wurde und Unterschiede im akzessorischem Gengehalt zurücklässt. Hierbei würde die Analyse des akzessorischen Genoms wiederum, im Gegensatz zu dem des Kerngenoms, interessante Einblicke in die Populationsstruktur geben[162]. Die Anwendung eines Grenzwertes, der auf der Mutationsrate innerhalb eines Patienten beruht, ist somit für Isolate, deren Isolationsdaten weit auseinander liegen, nicht relevant[27].

Prinzipiell zeigen die Ergebnisse des hoch diversen Datensatzes, dass der oftmals angewendete Grenzwert für nahe genomische Verwandtschaften eher eine Richtlinie als eine absolute Grenze darstellt und aufgrund der hier genannten Argumente mit einer gewissen Flexibilität betrachtet werden sollte[162]. Zudem können auch Ausbrüche auftreten, die nicht klonal sind. Ein Grenzwert zum Aus- bzw. Einschluss von Infektionsfällen in einen Ausbruch stellt demnach eine biologisch fragwürdige Maßnahme dar[107]. Die Bestimmung eines passenden Grenzwertes wurde auch schon von Ruan et al. als problematisch angesehen[165]. Neben der

Aussage, dass für jede bakterielle Spezies ein individueller Grenzwert festgelegt werden muss, wurde auch hier angedeutet, dass eine Bestimmung eher für jeden Ausbruch individuell vorgenommen werden sollte. Dies mag für die durch Surveillance detektierten Ausbrüche möglich sein und kann nach einer gewissen Ansammlung an Genomen, die zu einem Ausbruch gehören, individuell für diesen berechnet werden. Allerdings setzt dieser Vorgang eben genau die initiale Erkennung eines Ausbruchs voraus und lässt keine umgekehrte genomische Epidemiologie zu. Das Anwenden eines Grenzwertes zur Detektion von nahen genomischen Verwandtschaften ist demnach zwar berechtigt, sollte aber immer als Hypothese für eine mögliche Transmission und nicht als endgültiger Beweis angesehen werden[107].

4.2.4 Aussicht für weitere Analysen der Populationsstruktur von *C. difficile*

Die in dieser Arbeit vorgenommene Typisierung von *C. difficile* Isolaten anhand der Clusterung ihrer Genome in HC150 Cluster deutete ein länderübergreifendes und teilweise über Jahrzehnte andauerndes Vorkommen vieler bisher nicht umfassend untersuchter endemischer Stämme an. Zudem bildeten sich neben schon bekannten und tiefer erforschten Stämmen phylogenetisch nah verwandte neue Cluster beziehungsweise Singletons, die möglicherweise ähnliche phänotypische Merkmale aufzeigen und sich zukünftig zu pandemischen oder endemischen Stämmen ausbilden könnten. Die Ausbreitung dieser Stämme wäre durch regelmäßiges untersuchen der HC150 Cluster früh erkennbar.

Des Weiteren kann, wie auch in Kapitel 4.1.1 diskutiert, mit Hilfe der HC150 Cluster eine passende Referenzsequenz für die SNP-Analyse ermittelt werden. Dafür müsste allerdings sicher gestellt sein, dass jedes HC150 Cluster ein komplett sequenziertes Genom enthält. Geeignete Einträge müssten in der Datenbank bestimmt und, wenn nötig, die Besitzer dieses Stammes kontaktiert werden um eine Genomrekonstruktion vornehmen zu können. Für den Großteil der HC150 Cluster könnte dies allerdings auch ohne Fremddaten erfolgen, da die im Laufe dieser Arbeit hochgeladenen Stämme eine große Diversität besitzen und am Leibniz-Institut DSMZ verfügbar sind.

Mit Hilfe des Werkzeugs GrapeTree könnte in EnteroBase zudem eine Übersicht über die Verteilung phänotypischer Eigenschaften in der *C. difficile* Population gewonnen werden, indem sich phylogenetische Bäume nach unterschiedlichen Metadaten, sowie nach Allelnummern der wgMLST und cgMLST Schemata einfärben lassen. Für PCR Ribotypen wurde schon gezeigt, dass sie mit unterschiedlichen DNA-Sequenzen für verschiedene Gene korrelieren[174]. Um mögliche Muster phänotypischer Eigenschaften abbilden zu können, müssten diese allerdings auf eine Punktmutation zurückzuführen sein, die durch die Loci in einem der zur Verfügung stehenden Schemata erfasst ist. Hierbei wäre es natürlich von großem Interesse die Verteilung von Antibiotikaresistenzen darzustellen und näher zu untersuchen.

In der vorliegenden Arbeit wurde die cgMLST-Analyse erfolgreich zur Ausbruchsdetektion verwendet. Die Anwendung des anhand der Beobachtungen von Eyre et al. etablierten Grenzwertes von ≤ 2 SNPs für mögliche Transmissionsketten zeigte im weiteren Verlauf der Arbeit allerdings auch Limitierungen auf, besonders bei der Analyse von unabhängig voneinander isolierten Isolaten. Hier schienen sich die Isolate trotz naher Verwandtschaft im Kerngenom deutlich in ihrem akzessorischen Gengehalt zu unterscheiden. Wie genau sich Änderungen im akzessorischen Genom auf die Evolution von *C. difficile* auswirken und ob zwei Isolate mit abweichendem akzessorischem Gengehalt und identischen Kerngenomen trotzdem zu einer Transmissionskette gezählt werden können, muss noch näher untersucht werden. Dafür wäre es interessant eine molekulare Uhr für das akzessorische Genom zu berechnen, um ein mögliches zeitliches Signal der Mutationen zu verzeichnen. Diese Analyse könnte für Isolate aus ausgewählten HC150 Clustern durchgeführt werden, für die Jahres- und Länderinformationen verfügbar sind und deren Kerngenome eine nahe Verwandtschaft aufzeigen.

4.3 Fazit

Mit der hierarchischen Clusterung der Genome in HC150 Cluster können *C. difficile* Isolate zukünftig alternativ zur standardmäßig angewendeten PCR Ribotypisierung in endemische Stämme eingeteilt werden. Im Gegensatz zur PCR Ribotypisierung werden Isolate anhand der Typisierung durch Clusterung in HC150 Cluster (cgST Complexe; CC) einheitlich bezeichnet, sodass endemische Stämme in Zukunft weltweit miteinander verglichen werden können. Dies gilt auch für pandemische Stämme (HC10 Cluster) oder übergeordnete Populationslevel, die durch höhere hierarchische Cluster untersucht werden können.
Die Distanzwerte der cgMLST- und SNP-Analyse korrelieren stark miteinander. Der in der Literatur etablierte Wert von $\leq$2 SNPs zur Detektion von Transmissionsketten konnte somit auf die berechneten Kerngenom-Allelunterschiede in der cgMLST-Analyse angewendet werden. Dadurch konnten retrospektiv mögliche Transmissionwege in einem Netzwerk von Krankenhäusern aufgedeckt und rezidivierende CDI von Neuinfektionen unterschieden werden. Neben der Analyse von klinischen Isolaten konnte die cgMLST-Analyse zudem aufdecken, dass sich durch die Ausbringung von tierischem Dung auf landwirtschaftlich genutzte Flächen *C. difficile* Isolate aus dem Mist durch aufgewirbelte Staubpartikel ausbreiten können.
Allerdings deckte die Analyse der Isolate der gesamten Datenbank auf nahe genomische Verwandtschaften Limitierungen der cgMLST- wie auch der SNP-Analyse auf. Isolate, die sich sowohl in ihrem Herkunftsland als auch teilweise stark in ihrem Isolationszeitpunkt unterscheiden, zeigten trotz naher genomischer Verwandtschaft im Kerngenom verschiedene Gengehalte in akzessorischen Genomabschnitten. Dadurch war die Zuordnung dieser Isolate zu einem Ausbruchsklon eher als implausibel anzunehmen. Ein direkter Zusammenhang zweier Isolate durch eine Transmission sollte erst dann geschlussfolgert werden, wenn die genomischen Distanzen sowohl im Kern- als auch im akzessorischen Genom mit den zugehörigen epidemiologischen Daten ein plausibles Bild abgeben.

Diese Arbeit demonstrierte erfolgreich den umfangreichen Nutzen von EnteroBase und das anhand dieser Plattform in Zukunft Typisierungen und Ausbruchsanalysen von Wissenschaftlern ohne bioinformatischem Hintergrund in einem globalen Kontext einheitlich durchgeführt werden können. Zudem zeigte sich das Konzept der umgekehrten genomischen Epidemiologie als hilfreiches Mittel bei der Aufdeckung von Ausbrüchen oder rezidivierender CDI, sodass diese in Zukunft routinemäßig durchgeführt werden sollte um die epidemiologische Surveillance zu lenken und gezielte Hygienemaßnahmen treffen zu können.

Anhang A

Isolatelisten

Tabelle A.1: Isolateliste des Datensatzes von Patienten mit rezidivierender CDI (Kapitel 2.2.1). Die zugehörigen Sequenzen wurden im Europäischen Nukleotid Archiv unter der *Study Accession number* PRJEB33768 hinterlegt.

Isolatebezeichnung	Land	Isolationsquelle	Isolationsdatum	Patient
CD-15-00985	Deutschland	Human	20.11.2014	G
CD-15-00984	Deutschland	Human	20.11.2014	G
CD-15-00982	Deutschland	Human	20.11.2014	G
CD-15-00981	Deutschland	Human	20.11.2014	G
CD-15-00980	Deutschland	Human	20.06.2014	G
CD-15-00979	Deutschland	Human	20.06.2014	G
CD-15-00978	Deutschland	Human	20.06.2014	G
CD-15-00977	Deutschland	Human	20.06.2014	G
CD-15-00976	Deutschland	Human	20.06.2014	G
CD-15-00975	Deutschland	Human	20.06.2014	G
CD-15-00974	Deutschland	Human	20.06.2014	G
CD-15-00973	Deutschland	Human	20.06.2014	G
CD-15-00972	Deutschland	Human	20.06.2014	G
CD-15-00970	Deutschland	Human	20.06.2014	G
CD-15-00969	Deutschland	Human	20.06.2014	G
CD-15-00968	Deutschland	Human	20.06.2014	G
CD-15-00591	Deutschland	Human	11.11.2014	F
CD-15-00590	Deutschland	Human	11.11.2014	F
CD-15-00589	Deutschland	Human	11.11.2014	F
CD-15-00588	Deutschland	Human	11.11.2014	F
CD-15-00587	Deutschland	Human	11.11.2014	F
CD-15-00586	Deutschland	Human	11.11.2014	F
CD-15-00580	Deutschland	Human	11.11.2014	F
CD-15-00576	Deutschland	Human	11.11.2014	F
CD-15-00575	Deutschland	Human	11.11.2014	F
CD-15-00574	Deutschland	Human	11.11.2014	F
CD-15-00573	Deutschland	Human	11.11.2014	F
CD-15-00572	Deutschland	Human	11.11.2014	F
CD-15-00571	Deutschland	Human	11.11.2014	F
CD-15-00570	Deutschland	Human	11.11.2014	F

CD-15-00569	Deutschland	Human	11.11.2014	F
CD-15-00568	Deutschland	Human	11.11.2014	F
CD-15-00566	Deutschland	Human	11.11.2014	F
CD-15-00564	Deutschland	Human	11.11.2014	F
CD-15-00563	Deutschland	Human	11.11.2014	F
CD-15-00562	Deutschland	Human	11.11.2014	F
CD-15-00561	Deutschland	Human	11.11.2014	F
CD-15-00560	Deutschland	Human	11.11.2014	F
CD-15-00559	Deutschland	Human	11.11.2014	F
CD-15-00558	Deutschland	Human	11.11.2014	F
CD-15-00557	Deutschland	Human	11.11.2014	F
CD-15-00556	Deutschland	Human	11.11.2014	F
CD-15-00555	Deutschland	Human	11.11.2014	F
CD-15-00554	Deutschland	Human	11.11.2014	F
CD-15-00553	Deutschland	Human	11.11.2014	F
CD-15-00552	Deutschland	Human	11.11.2014	F
CD-15-00551	Deutschland	Human	11.11.2014	F
CD-15-00549	Deutschland	Human	11.11.2014	F
CD-15-00548	Deutschland	Human	11.11.2014	F
CD-15-00547	Deutschland	Human	11.11.2014	F
CD-15-00546	Deutschland	Human	11.11.2014	F
CD-15-00545	Deutschland	Human	11.11.2014	F
CD-15-00383	Deutschland	Human	22.07.2014	F
CD-15-00382	Deutschland	Human	22.07.2014	F
CD-15-00381	Deutschland	Human	22.07.2014	F
CD-15-00380	Deutschland	Human	22.07.2014	F
CD-15-00379	Deutschland	Human	22.07.2014	F
CD-15-00378	Deutschland	Human	22.07.2014	F
CD-15-00377	Deutschland	Human	22.07.2014	F
CD-15-00376	Deutschland	Human	22.07.2014	F
CD-15-00375	Deutschland	Human	22.07.2014	F
CD-15-00374	Deutschland	Human	22.07.2014	F
CD-15-00373	Deutschland	Human	22.07.2014	F
CD-15-00372	Deutschland	Human	22.07.2014	F
CD-15-00371	Deutschland	Human	22.07.2014	F
CD-15-00370	Deutschland	Human	22.07.2014	F
CD-15-00369	Deutschland	Human	22.07.2014	F
CD-15-00368	Deutschland	Human	22.07.2014	F
CD-15-00367	Deutschland	Human	22.07.2014	F
CD-15-00366	Deutschland	Human	22.07.2014	F
CD-15-00365	Deutschland	Human	22.07.2014	F
CD-15-00364	Deutschland	Human	22.07.2014	F
CD-15-00363	Deutschland	Human	22.07.2014	F
CD-15-00362	Deutschland	Human	22.07.2014	F
CD-15-00361	Deutschland	Human	22.07.2014	F
CD-15-00360	Deutschland	Human	22.07.2014	F
CD-15-00359	Deutschland	Human	22.07.2014	F

CD-15-00358	Deutschland	Human	22.07.2014	F
CD-15-00357	Deutschland	Human	22.07.2014	F
CD-15-00356	Deutschland	Human	22.07.2014	F
CD-15-00355	Deutschland	Human	22.07.2014	F
CD-15-00354	Deutschland	Human	22.07.2014	F
CD-15-00353	Deutschland	Human	22.07.2014	F
CD-15-00352	Deutschland	Human	22.07.2014	F
CD-15-00351	Deutschland	Human	22.07.2014	F
CD-15-00284	Deutschland	Human	21.01.2015	D
CD-15-00283	Deutschland	Human	21.01.2015	D
CD-15-00282	Deutschland	Human	21.01.2015	D
CD-15-00281	Deutschland	Human	21.01.2015	D
CD-15-00253	Deutschland	Human	17.07.2014	E
CD-15-00251	Deutschland	Human	17.07.2014	E
CD-15-00248	Deutschland	Human	17.07.2014	E
CD-15-00247	Deutschland	Human	17.07.2014	E
CD-15-00246	Deutschland	Human	17.07.2014	E
CD-15-00245	Deutschland	Human	17.07.2014	E
CD-15-00243	Deutschland	Human	17.07.2014	E
CD-15-00242	Deutschland	Human	17.07.2014	E
CD-15-00241	Deutschland	Human	17.07.2014	E
CD-15-00237	Deutschland	Human	17.07.2014	E
CD-15-00236	Deutschland	Human	17.07.2014	E
CD-15-00233	Deutschland	Human	17.07.2014	E
CD-15-00232	Deutschland	Human	17.07.2014	E
CD-15-00231	Deutschland	Human	17.07.2014	E
CD-15-00230	Deutschland	Human	17.07.2014	E
CD-15-00229	Deutschland	Human	17.07.2014	E
CD-15-00228	Deutschland	Human	17.07.2014	E
CD-15-00227	Deutschland	Human	17.07.2014	E
CD-15-00226	Deutschland	Human	17.07.2014	E
CD-15-00225	Deutschland	Human	17.07.2014	E
CD-15-00223	Deutschland	Human	17.07.2014	E
CD-15-00222	Deutschland	Human	17.07.2014	E
CD-15-00221	Deutschland	Human	17.07.2014	E
CD-15-00220	Deutschland	Human	17.07.2014	E
CD-15-00219	Deutschland	Human	17.07.2014	E
CD-15-00218	Deutschland	Human	17.07.2014	E
CD-15-00217	Deutschland	Human	17.07.2014	E
CD-15-00216	Deutschland	Human	17.07.2014	E
CD-15-00215	Deutschland	Human	17.07.2014	E
CD-15-00214	Deutschland	Human	17.07.2014	E
CD-15-00213	Deutschland	Human	17.07.2014	E
CD-15-00212	Deutschland	Human	17.07.2014	E
CD-15-00211	Deutschland	Human	17.07.2014	E
CD-15-00210	Deutschland	Human	17.07.2014	E
CD-15-00209	Deutschland	Human	17.07.2014	E

CD-15-00208	Deutschland	Human	17.07.2014	E
CD-15-00207	Deutschland	Human	17.07.2014	E
CD-15-00206	Deutschland	Human	28.04.2014	E
CD-15-00205	Deutschland	Human	28.04.2014	E
CD-15-00204	Deutschland	Human	28.04.2014	E
CD-15-00203	Deutschland	Human	28.04.2014	E
CD-15-00202	Deutschland	Human	28.04.2014	E
CD-15-00201	Deutschland	Human	28.04.2014	E
CD-15-00200	Deutschland	Human	28.04.2014	E
CD-15-00199	Deutschland	Human	28.04.2014	E
CD-15-00198	Deutschland	Human	28.04.2014	E
CD-15-00197	Deutschland	Human	28.04.2014	E
CD-15-00196	Deutschland	Human	28.04.2014	E
CD-15-00195	Deutschland	Human	28.04.2014	E
CD-15-00193	Deutschland	Human	28.04.2014	E
CD-15-00192	Deutschland	Human	28.04.2014	E
CD-15-00191	Deutschland	Human	28.04.2014	E
CD-15-00190	Deutschland	Human	28.04.2014	E
CD-15-00189	Deutschland	Human	28.04.2014	E
CD-15-00188	Deutschland	Human	28.04.2014	E
CD-15-00187	Deutschland	Human	28.04.2014	E
CD-15-00186	Deutschland	Human	28.04.2014	E
CD-15-00184	Deutschland	Human	28.04.2014	E
CD-15-00181	Deutschland	Human	28.04.2014	E
CD-15-00180	Deutschland	Human	28.04.2014	E
CD-15-00179	Deutschland	Human	28.04.2014	E
CD-15-00177	Deutschland	Human	28.04.2014	E
CD-15-00175	Deutschland	Human	28.04.2014	E
CD-15-00174	Deutschland	Human	28.04.2014	E
CD-15-00173	Deutschland	Human	28.04.2014	E
CD-15-00172	Deutschland	Human	21.01.2015	D
CD-15-00170	Deutschland	Human	21.01.2015	D
CD-15-00169	Deutschland	Human	21.01.2015	D
CD-15-00168	Deutschland	Human	21.01.2015	D
CD-15-00167	Deutschland	Human	21.01.2015	D
CD-15-00166	Deutschland	Human	21.01.2015	D
CD-15-00165	Deutschland	Human	21.01.2015	D
CD-15-00163	Deutschland	Human	21.01.2015	D
CD-15-00161	Deutschland	Human	21.01.2015	D
CD-15-00160	Deutschland	Human	21.01.2015	D
CD-15-00159	Deutschland	Human	27.08.2014	D
CD-15-00158	Deutschland	Human	27.08.2014	D
CD-15-00157	Deutschland	Human	27.08.2014	D
CD-15-00156	Deutschland	Human	27.08.2014	D
CD-15-00155	Deutschland	Human	27.08.2014	D
CD-15-00152	Deutschland	Human	27.08.2014	D
CD-15-00151	Deutschland	Human	27.08.2014	D

CD-15-00150	Deutschland	Human	27.08.2014	D
CD-15-00149	Deutschland	Human	27.08.2014	D
CD-15-00146	Deutschland	Human	27.08.2014	D
CD-15-00145	Deutschland	Human	27.08.2014	D
CD-15-00144	Deutschland	Human	27.08.2014	D
CD-15-00143	Deutschland	Human	27.08.2014	D
CD-15-00142	Deutschland	Human	27.08.2014	D
CD-15-00141	Deutschland	Human	27.08.2014	D

Tabelle A.2: Isolateliste des Datensatzes aus einem Netzwerk von Krankenhäusern (Kapitel 2.2.1). Die Isolate wurden aus humanen Proben aus Deutschland isoliert und die zugehörige Krankenhaus- bzw. Stationsinformation wurde anonymisiert. Die zugehörigen Sequenzen wurden im Europäischen Nukleotid Archiv unter der*Study Accession number* PRJEB33779 hinterlegt.

Isolatebezeichnung	Kranken-haus	HC2	Isolations-station	Isolations-datum	weitere Station	Datum	weitere Station	Datum	weitere Station	Datum	Kommentar	PCR Ribotyp
CD-15-00632	5	4790		2013								014
CD-15-00633	5	76		2013								001
CD-15-00634	5	4791		2013								241
CD-15-00636	5	4399		2013								001
CD-15-00638	5	1131	5_10	25.11.2013								027
CD-15-00639	5	109	5_7	13.09.2013	6_1							001
CD-15-00640	5	1251	5_11	09.10.2013	3_2							001
CD-15-00641	5	4794		2013								014
CD-15-00642	5	4796		2013								614
CD-15-00643	5	4798		2013								078
CD-15-00644	5	76	5_17	23.10.2013	5_24		5_22		5_21			001
CD-15-00645	5	4795		2013								014
CD-15-00646	5	76	5_5	31.10.2013								001
CD-15-00647	5	4592		2013								220
CD-15-00648	5	76	5_26	2013	5_7		5_26				Rezidiv mit CD-15-00728	001
CD-15-00649	5	76	5_11	12.12.2013	5_21							001
CD-15-00650	5	1289		2013								012
CD-15-00651	5	4797		2013								012
CD-15-00652	3	1287		2013								001
CD-15-00653	3	1251	3_2	07.09.2013	3_6							001
CD-15-00655	3	2	3_3	25.09.2013								002
CD-15-00656	3	1251	3_2	10.10.2013								001
CD-15-00657	3	1267	3_2	10.10.2013								n. d.
CD-15-00658	3	70	3_2	06.11.2013								001
CD-15-00660	3	76	3_11	15.11.2013	2_1	05.11.2013						001
CD-15-00661	3	4534		2013								001
CD-15-00662	3	76	3_10	23.11.2013								n. d.

Isolatebezeichnung	Kranken-haus	HC2	Isolations-station	Isolations-datum	weitere Station	Datum	weitere Station	Datum	weitere Station	Datum	Kommentar	PCR Ribotyp
CD-15-00663	3	85	3_5	25.11.2013	3_6							027
CD-15-00664	3	76	3_4	25.11.2013								001
CD-15-00665	3	70	3_2	03.12.2013	3_7		5_12					001
CD-15-00666	3	80		2013								002
CD-15-00668	3	76	3_9	13.12.2013								001
CD-15-00669	3	76	3_2	19.12.2013								001
CD-15-00672	1	1277		2013								081
CD-15-00673	1	1243	1_1	19.11.2013								236
CD-15-00674	1	1276		2013								020
CD-15-00675	1	1352		2013								051
CD-15-00676	1	1275		2013								620
CD-15-00677	1	1279		2014								n. d.
CD-15-00679	1	1273		2014								014
CD-15-00681	1	4805		2014								014
CD-15-00682	1	4536		2014								001
CD-15-00683	1	1280		2014								005
CD-15-00684	1	1272		2014								027
CD-15-00685	1	3262		2014								078
CD-15-00686	1	4806		2014								126
CD-15-00688	1	1271		2014								241
CD-15-00689	1	76	1_1	24.06.2014								001
CD-15-00690	1	76	1_1	19.04.2014								001
CD-15-00692	2	1267	2_3	01.09.2013	2_4							n. d.
CD-15-00693	2	1267	2_2	07.09.2013								n. d.
CD-15-00694	2	4600		2013								023
CD-15-00695	2	4593		2013								020
CD-15-00696	2	76	2_5	24.10.2013								001
CD-15-00697	2	76	2_5	01.11.2013								001
CD-15-00698	2	1265		2013								014
CD-15-00699	2	1264		2013								005
CD-15-00700	2	76	2_5	05.12.2013	3_2	22.11.2013	3_7	26.11.2013	3_2	10.12.2013		001

Isolatebezeichnung	Kranken-haus	HC2	Isolations-station	Isolations-datum	weitere Station	Datum	weitere Station	Datum	weitere Station	Datum	Kommentar	PCR Ribotyp
CD-15-00701	2	76	2_5	14.12.2013	2_1	15.11.2013						001
CD-15-00702	2	76	2_5	18.12.2013	3_7	16.11.2013						n. d.
CD-15-00703	2	1262		2013								023
CD-15-00704	2	1261		2014								721
CD-15-00705	2	1260		2014								708
CD-15-00706	2	4594		2014								014
CD-15-00707	2	76	2_5	15.04.2014								001
CD-15-00708	2	76	2_2	16.04.2014								001
CD-15-00709	2	1256		2014								039
CD-15-00710	2	1258		2014								087
CD-15-00711	2	76	2_5	22.04.2014								241
CD-15-00714	5	5164		2013								559
CD-15-00715	3	76	3_9	16.12.2013	3_2							001
CD-15-00716	3	79		2013								010
CD-15-00717	3	1252		2013								n. d.
CD-15-00718	6	1251	6_2	31.08.2013	5_8							001
CD-15-00720	6	76	6_1	28.10.2013	5_25							001
CD-15-00721	6	1251	6_1	28.10.2013	6_2							001
CD-15-00722	6	76	6_4	28.10.2013	6_1							001
CD-15-00723	6	1245		2013								002
CD-15-00724	6	4599		2013								014
CD-15-00725	6	1240		2013								002
CD-15-00726	6	76	6_1	26.12.2013								001
CD-15-00727	6	109	6_2	14.02.2014	5_16		5_7		5_17			001
CD-15-00728	6	76	6_1	01.03.2014							Rezidiv mit CD-15-00648	001
CD-15-00729	6	1248		2014								722
CD-15-00730	6	1234		2014								078
CD-15-00731	6	1233		2014								003
CD-15-00732	6	1236		2014								014
CD-15-00733	6	1232	6_1	13.05.2014	5_1		5_2					001
CD-15-00734	6	109	6_1	05.06.2014	6_2							001

Isolatebezeichnung	Kranken-haus	HC2	Isolations-station	Isolations-datum	weitere Station	Datum	weitere Station	Datum	weitere Station	Datum	Kommentar	PCR Ribotyp
CD-15-00735	6	1244		2014								002
CD-15-00736	6	1243	6_2	25.07.2014								236
CD-15-00737	6	1247		2014								002
CD-15-00738	6	1242	6_3	13.06.2014								001
CD-15-00855	5	4816		2014								062
CD-15-00856	5	479	5_14	03.09.2014								002
CD-15-00857	5	1131	5_18	04.09.2014								027
CD-15-00858	5	1294		2014								446
CD-15-00859	5	479	5_19	10.09.2014								002
CD-15-00860	5	1127	5_13	11.09.2014								001
CD-15-00861	5	4809		2014								n. d.
CD-15-00862	5	4808	5_8	22.09.2014								449
CD-15-00863	5	1127	5_13	23.09.2014								001
CD-15-00864	5	1349		2014								014
CD-15-00865	5	1126		2014								027
CD-15-00866	5	1127	5_8	10.10.2014	5_15		5_13					001
CD-15-00867	5	4814		2014								031
CD-15-00868	5	4818		2014								001
CD-15-00869	5	1127	5_13	12.10.2014								001
CD-15-00870	5	4820		2014								020
CD-15-00871	5	4824		2014								078
CD-15-00872	5	4808	5_2	18.10.2014	5_34		5_24					449
CD-15-00873	5	4815		2014								078
CD-15-00874	5	1127	5_29	19.10.2014								001
CD-15-00875	5	1127	5_8	19.10.2014	5_15	15.09.2014						001
CD-15-00876	2	1210	2_5	31.08.2014								023
CD-15-00877	2	1298		2014								078
CD-15-00878	2	4819		2014								014
CD-15-00879	2	4829		2014								039
CD-15-00880	2	4823	2_8	11.09.2014	3_7							014
CD-15-00881	2	76	2_9	11.09.2014								001

Isolatebezeichnung	Kranken-haus	HC2	Isolations-station	Isolations-datum	weitere Station	Datum	weitere Station	Datum	weitere Station	Datum	Kommentar	PCR Ribotyp
CD-15-00882	2	176		2014								720
CD-15-00883	2	1293		2014								003
CD-15-00884	2	4834		2014								012
CD-15-00885	2	172		2014								001
CD-15-00887	2	1348		2014								012
CD-15-00888	2	1355		2014								719
CD-15-00889	2	1131	2_6	07.10.2014	2_7							027
CD-15-00890	2	4823	2_5	03.10.2014								014
CD-15-00891	2	4825		2014								023
CD-15-00892	2	1227		2014								039
CD-15-00893	2	4828		2014								078
CD-15-00894	2	4831		2014								668
CD-15-00895	2	1127	2_1	26.10.2014	2_9	04.08.2014	5_30	04.10.2014	5_13			001
CD-15-00920	1	1225	1_3	10.10.2014	5_20		1_1					241
CD-15-00921	1	1224		2014								078
CD-15-00922	1	1222		2014								014
CD-15-00923	1	1226		2014								023
CD-15-00924	1	186		2014								014
CD-15-00925	1	1208	1_1	31.10.2014								001
CD-15-00926	1	1220		2014								666
CD-15-00927	1	1223		2014								006
CD-15-00928	1	1219		2014								n. d.
CD-15-00929	1	1218		2015								451
CD-15-00930	1	1217		2015								003
CD-15-00931	1	1206	1_2	26.01.2015								027
CD-15-00932	1	1122		2015								023
CD-15-00933	1	1215		2015								225
CD-15-00934	1	1225	1_1	10.03.2015								241
CD-15-00935	1	76	1_2	14.04.2015								001
CD-15-00936	1	76	1_2	14.04.2015	1_1	01.03.2015						001
CD-15-00937	1	76	1_1	13.04.2015								001

Isolatebezeichnung	Kranken-haus	HC2	Isolations-station	Isolations-datum	weitere Station	Datum	weitere Station	Datum	weitere Station	Datum	Kommentar	PCR Ribotyp
CD-15-00938	6	1209		2014								046
CD-15-00939	6	171		2014								020
CD-15-00940	6	1210	6_2	14.11.2014	6_1		5_3					023
CD-15-00941	6	76	6_2	08.12.2014	5_33	01.11.2014						001
CD-15-00942	6	1208	6_2	17.12.2014								001
CD-15-00943	6	76	6_1	15.01.2015								001
CD-15-00944	6	1207		2015								084
CD-15-00945	6	1206	6_4	20.01.2015	5_4							027
CD-15-00946	6	76	6_3	20.01.2015								456
CD-15-00947	6	1205		2015								n. d.
CD-15-00948	6	76	6_2	20.02.2015								001
CD-15-00949	3	76	3_7	24.06.2014								001
CD-15-00950	3	76	3_2	05.09.2014	2_5	01.07.2014						001
CD-15-00951	3	1214		2014								n. d.
CD-15-00952	3	76	3_13	07.09.2014								001
CD-15-00953	3	76	3_2	09.09.2014								001
CD-15-00954	3	70	3_1	09.09.2014								001
CD-15-00955	3	76	3_6	09.09.2014								001
CD-15-00956	3	76	3_14	15.09.2014								001
CD-15-00957	3	1198		2014								049
CD-15-00958	3	1202		2014								010
CD-15-00959	3	76	3_9	08.10.2014								001
CD-15-00960	3	1347		2014								002
CD-15-00961	3	282		2014								020
CD-15-00962	3	76	3_6	05.11.2014								001
CD-15-00963	3	76	3_6	07.11.2014	2_5	01.10.2014						001
CD-15-00964	3	76	3_6	10.11.2014								001
CD-15-00966	3	1195		2014								563
CD-15-00967	3	1192		2014								029
CD-16-00083	3	4608		2015								014
CD-16-00084	3	4565		2015								453

CD-16-00085	3	76	3_3	30.09.2015	3_8							001
CD-16-00086	3	1131	3_5	08.10.2015								027
CD-16-00087	3	4561		2015								001
CD-16-00088	3	310		2014								027
CD-16-00089	3	4562		2015								002
CD-16-00090	3	4574		2015								020
CD-16-00091	3	4570		2015								547
CD-16-00092	3	4564		2015								446
CD-16-00093	3	4839		2015								001
CD-16-00094	3	4577		2015								014
CD-16-00095	3	76	3_12	29.11.2015								001
CD-16-00096	3	76	3_2	05.12.2015	5_1	01.11.2015						001
CD-16-00097	3	4837		2015								014
CD-16-00098	3	1225	3_2	08.12.2015								001
CD-16-00099	3	4572		2015								005
CD-16-00100	3	4567		2015								106
CD-16-00101	3	85	3_4	15.01.2016								027
CD-16-00102	1	76	1_1	02.09.2015	1_3							001
CD-16-00103	1	4571		2015								005
CD-16-00104	1	4579		2015								001
CD-16-00105	1	4575		2015								005
CD-16-00106	1	76	1_2	12.10.2015	5_7							001
CD-16-00107	1	4578		2015								001
CD-16-00108	1	76	1_4	05.10.2015								001
CD-16-00109	1	76	1_4	16.10.2015	1_2	08.10.2015						001
CD-16-00110	1	76	1_2	21.10.2015	1_1	11.08.2015	23/09/15					001
CD-16-00111	1	4583		2015								014
CD-16-00112	1	491	1_1	11.11.2015	1_4							023
CD-16-00113	1	491	1_1	11.11.2015								023
CD-16-00114	1	4405		2015								049
CD-16-00115	1	4404		2016								027
CD-16-00116	1	76	1_1	14.01.2016								001

Isolatebezeichnung	Kranken-haus	HC2	Isolations-station	Isolations-datum	weitere Station	Datum	weitere Station	Datum	weitere Station	Datum	Kommentar	PCR Ribotyp
CD-16-00117	2	4406		2015								062
CD-16-00118	2	4407		2015								020
CD-16-00119	2	1206	2_11	2015								027
CD-16-00120	2	4408		2015								220
CD-16-00121	2	4409		2015								014
CD-16-00122	2	76	2_9	17.09.2015	2_10	01.04.2015	3_3				Rezidiv mit CD-16-00138	001
CD-16-00123	2	4410		2015								003
CD-16-00124	2	4411		2015								012
CD-16-00125	2	76	2_5	25.09.2015								001
CD-16-00126	2	4414		2015								002
CD-16-00127	2	76	2_5	29.09.2015								001
CD-16-00128	2	4415		2015							Rezidiv mit CD-16-00129	078
CD-16-00129	2	4415		2015							Rezidiv mit CD-16-00128	078
CD-16-00130	2	5074		2015								014
CD-16-00131	2	4417		2015								003
CD-16-00133	2	5072		2015								n. d.
CD-16-00134	2	1206	2_12	2015								027
CD-16-00136	2	4420		2015								003
CD-16-00137	2	1206	2_11	2015								027
CD-16-00138	2	76	2_9	17.11.2015							Rezidiv mit CD-16-00122	001
CD-16-00139	4	5078		2015								023
CD-16-00140	4	4604		2015								003
CD-16-00141	4	5076		2015								001
CD-16-00142	4	2418		2015								012
CD-16-00143	4	4603		2015								015
CD-16-00144	4	76	4_3	14.10.2015	6_1	2013						001
CD-16-00145	4	4425		2015								027
CD-16-00146	4	4607		2015								027
CD-16-00147	4	4426		2015								020
CD-16-00148	4	4427		2015								078
CD-16-00149	4	4428		2015								212

Isolatebezeichnung	Kranken-haus	HC2	Isolations-station	Isolations-datum	weitere Station	Datum	weitere Station	Datum	weitere Station	Datum	Kommentar	PCR Ribotyp
CD-16-00150	4	76	4_4	25.11.2015								001
CD-16-00151	4	4431	4_3	03.12.2015								001
CD-16-00152	4	4430		2015								027
CD-16-00153	4	1242	4_1	04.12.2015								001
CD-16-00154	4	1127	4_1	06.12.2015								001
CD-16-00155	4	5088		2015								014
CD-16-00156	4	4436		2015								651
CD-16-00157	4	4431	4_2	31.12.2015								001
CD-16-00158	4	1131	4_2	17.01.2016								027
CD-16-00159	5	4439		2015								005
CD-16-00160	5	4438		2015								002
CD-16-00161	5	1232	5_5	10.09.2015								001
CD-16-00162	5	76	5_28	10.09.2015	5_28	13.08.2015	5_15	20.08.2015				001
CD-16-00163	5	4441		2015								014
CD-16-00164	5	1243	5_6	12.09.2015								236
CD-16-00165	5	4444		2015								005
CD-16-00166	5	1	5_23	14.09.2015								078
CD-16-00167	5	4445		2015								049
CD-16-00168	5	1178		2015								629
CD-16-00169	5	76	5_15	16.09.2015								001
CD-16-00170	5	4447		2015								n. d.
CD-16-00171	5	76	5_31	01.10.2015	5_15		6_2	2014	6_3	2014		001
CD-16-00173	5	1	5_5	06.10.2015	5_21		5_22					078
CD-16-00174	5	4451		2015								084
CD-16-00175	5	4452		2015								009
CD-16-00176	5	4453		2015								070
CD-16-00177	5	4455		2015								014
CD-16-00178	5	76	5_27	14.10.2015	5_5		5_21					001
CD-16-00179	5	5087		2015								027
CD-16-00180	5	4456		2015								001
CD-16-00181	3	1206	3_10	2016	2_13							027

Isolatebezeichnung	Kranken-haus	HC2	Isolations-station	Isolations-datum	weitere Station	Datum	weitere Station	Datum	weitere Station	Datum	Kommentar	PCR Ribotyp
CD-16-00182	3	4587		2016								014
CD-16-00183	3	4585		2016								014
CD-16-00184	3	1208	3_2	29.01.2016								001
CD-16-00185	3	1	3_8	07.02.2016	3_5							078
CD-16-00186	1	4457		2015								037
CD-16-00187	1	4589		2016								027
CD-16-00188	1	76	1_4	03.03.2016								001
CD-16-00189	1	4459		2016								n. d.
CD-16-00190	1	76	1_1	19.03.2016	1_3							001
CD-16-00192	1	76	1_1	28.03.2016	5_18	01.06.2015	3_9		5_32	01.01.2016		001
CD-16-00193	1	76	1_3	05.04.2016								001
CD-16-00194	1	4461		2016								534
CD-16-00195	1	4462		2016								002
CD-16-00196	3	2	3_3	10.12.2013	5_4		5_9					n. d.

Tabelle A.4: Liste der Isolate mit PCR Ribotypen Information (Kapitel 2.2.1). Die zugehörigen Sequenzen wurden jeweils im Europäischen Nukleotid Archiv unter der *Study Accession number* hinterlegt.

Isolatebezeichnung	Land	Isolations-quelle	Isolations-datum	PCR Ribotyp	Referenz	Study Accession number
CD-16-00537	Deutschland	Human	2010	005	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00536	Deutschland	Human	2010	020	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00535	Deutschland	Human	2010	049	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00534	Deutschland	Human	2010	027	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00533	Deutschland	Human	2010	157	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00531	Deutschland	Human	2010	070	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00530	Deutschland	Human	2010	023	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00529	Deutschland	Human	2010	054	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00528	Deutschland	Human	2010	150	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00527	Deutschland	Human	2010	126	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00526	Deutschland	Human	2010	009	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00525	Deutschland	Human	2010	050	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00523	Deutschland	Human	2010	001	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00522	Deutschland	Human	2009	078	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00521	Deutschland	Human	2009	158	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00520	Deutschland	Human	2009	056	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00517	Deutschland	Human	2009	103	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00515	Deutschland	Human	2009	078	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00514	Deutschland	Human	2009	045	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00513	Deutschland	Human	2009	063	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00512	Deutschland	Human	2009	072	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00511	Deutschland	Human	2009	106	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00510	Deutschland	Human	2009	001	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00509	Deutschland	Human	2009	001	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00505	Deutschland	Human	2008	039	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00504	Deutschland	Human	2008	083	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00503	Deutschland	Human	2008	081	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00502	Deutschland	Human	2008	090	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00500	Deutschland	Human	2008	149	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00498	Deutschland	Human	2008	035	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00496	Deutschland	Human	2008	001	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00495	Deutschland	Human	2008	001	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00494	Deutschland	Human	2008	156	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00493	Deutschland	Human	2008	010	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00491	Deutschland	Human	2008	094	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00490	Deutschland	Human	2007	044	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00489	Deutschland	Human	2007	044	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00440	Deutschland	Human	2009	066	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00439	Deutschland	Human	2009	014	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00438	Deutschland	Human	2009	015	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00436	Deutschland	Human	2008	117	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00434	Deutschland	Human	2008	087	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00433	Deutschland	Human	2009	078	Zaiß et al. 2010[119]	PRJEB33868

CD-16-00432	Deutschland	Human	2009	046	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00431	Deutschland	Human	2009	002	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00430	Deutschland	Human	2009	053	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00429	Deutschland	Human	2008	003	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00428	Deutschland	Human	2010	033	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00427	Deutschland	Human	2010	012	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00426	Deutschland	Human	2010	042	Zaiß et al. 2010[119]	PRJEB33868
CD-16-00425	Deutschland	Human	2010	011	Zaiß et al. 2010[119]	PRJEB33868
10-00472	Deutschland	Human	2010	n. d.	Zaiß et al. 2010[119]	PRJEB33868
10-00411	Deutschland	Human	2010	001	Zaiß et al. 2010[119]	PRJEB33868
10-00360	Deutschland	Human	2010	001	Zaiß et al. 2010[119]	PRJEB33868
10-00328	Deutschland	Human	2010	078	Zaiß et al. 2010[119]	PRJEB33868
10-00280	Deutschland	Human	2010	078	Zaiß et al. 2010[119]	PRJEB33868
10-00235	Deutschland	Human	2010	078	Zaiß et al. 2010[119]	PRJEB33868
10-00021	Deutschland	Human	2010	001	Zaiß et al. 2010[119]	PRJEB33868
09-00410	Deutschland	Human	2009	n. d.	Zaiß et al. 2010[119]	PRJEB33868
09-00388	Deutschland	Human	2009	001	Zaiß et al. 2010[119]	PRJEB33868
09-00365	Deutschland	Human	2009	001	Zaiß et al. 2010[119]	PRJEB33868
09-00326	Deutschland	Human	2009	001	Zaiß et al. 2010[119]	PRJEB33868
09-00212	Deutschland	Human	2009	078	Zaiß et al. 2010[119]	PRJEB33868
09-00199	Deutschland	Human	2009	078	Zaiß et al. 2010[119]	PRJEB33868
09-00129	Deutschland	Human	2009	078	Zaiß et al. 2010[119]	PRJEB33868
08-00454	Deutschland	Human	2008	078	Zaiß et al. 2010[119]	PRJEB33868
08-00430	Deutschland	Human	2008	078	Zaiß et al. 2010[119]	PRJEB33868
08-00389	Deutschland	Human	2008	n. d.	Zaiß et al. 2010[119]	PRJEB33868
08-00216	Deutschland	Human	2007	078	Zaiß et al. 2010[119]	PRJEB33868
07-00028	Deutschland	Human	2007	n. d.	Zaiß et al. 2010[119]	PRJEB33868
07-00003	UK	Human	2007	001	Zaiß et al. 2010[119]	PRJEB33868

CD-16-00081	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866
CD-16-00080	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866
CD-16-00079	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866
CD-16-00078	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866
CD-16-00077	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866
CD-16-00076	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866
CD-16-00075	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866
CD-16-00073	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866
CD-16-00072	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-16-00071	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-16-00070	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-16-00069	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-16-00068	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-16-00067	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-16-00066	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-16-00065	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-16-00063	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866
CD-16-00062	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866

CD-16-00061	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866
CD-16-00060	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866
CD-16-00058	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866
CD-16-00057	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866
CD-16-00056	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866
CD-16-00055	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866
CD-16-00054	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-16-00053	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-16-00052	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-16-00051	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-16-00050	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-16-00049	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-16-00048	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-16-00047	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-16-00046	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-16-00045	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-16-00044	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-16-00043	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-16-00042	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-16-00041	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-16-00040	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-16-00039	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-16-00038	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-16-00037	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-16-00036	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-16-00035	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-16-00034	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-16-00033	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-16-00032	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-16-00030	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-16-00029	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-16-00027	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-16-00026	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-16-00025	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-15-01046	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-15-01045	Deutschland	Human	NA	053	diese Arbeit	PRJEB33866
CD-15-01044	Deutschland	Human	NA	050	diese Arbeit	PRJEB33866
CD-15-01043	Deutschland	Human	NA	328	diese Arbeit	PRJEB33866
CD-15-01041	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866
CD-15-01039	Deutschland	Human	NA	078	diese Arbeit	PRJEB33866
CD-15-01037	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-15-01036	Deutschland	Human	NA	002	diese Arbeit	PRJEB33866
CD-15-01035	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-15-01034	Deutschland	Human	NA	n. d.	diese Arbeit	PRJEB33866
CD-15-01033	Deutschland	Human	NA	328	diese Arbeit	PRJEB33866
CD-15-01032	Deutschland	Human	NA	n. d.	diese Arbeit	PRJEB33866
CD-15-01031	Deutschland	Human	NA	003	diese Arbeit	PRJEB33866

CD-15-01030	Deutschland	Human	NA	026	diese Arbeit	PRJEB33866
CD-15-01029	Deutschland	Human	NA	n. d.	diese Arbeit	PRJEB33866
CD-15-01028	Deutschland	Human	NA	032	diese Arbeit	PRJEB33866
CD-15-01027	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-15-01026	Deutschland	Human	NA	073	diese Arbeit	PRJEB33866
CD-15-01025	Deutschland	Human	NA	010	diese Arbeit	PRJEB33866
CD-15-01024	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-15-01023	Deutschland	Human	NA	n. d.	diese Arbeit	PRJEB33866
CD-15-01022	Deutschland	Human	NA	226	diese Arbeit	PRJEB33866
CD-15-01021	Deutschland	Human	NA	018	diese Arbeit	PRJEB33866
CD-15-01020	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-15-01019	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-15-01017	Deutschland	Human	NA	001	diese Arbeit	PRJEB33866
CD-15-01016	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-15-01015	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-15-01014	Deutschland	Human	NA	n. d.	diese Arbeit	PRJEB33866
CD-15-01013	Deutschland	Human	NA	n. d.	diese Arbeit	PRJEB33866
CD-15-01012	Deutschland	Human	NA	n. d.	diese Arbeit	PRJEB33866
CD-15-01011	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866
CD-15-01010	Deutschland	Human	NA	n. d.	diese Arbeit	PRJEB33866
CD-15-01009	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866
CD-15-01008	Deutschland	Human	NA	017	diese Arbeit	PRJEB33866
CD-15-01007	Deutschland	Human	NA	024	diese Arbeit	PRJEB33866
CD-15-01006	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-15-01005	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-15-01004	Deutschland	Human	NA	106	diese Arbeit	PRJEB33866
CD-15-01003	Deutschland	Human	NA	053	diese Arbeit	PRJEB33866
CD-15-01002	Deutschland	Human	NA	018	diese Arbeit	PRJEB33866
CD-15-01001	Deutschland	Human	NA	002	diese Arbeit	PRJEB33866
CD-15-01000	Deutschland	Human	NA	073	diese Arbeit	PRJEB33866
CD-15-00999	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866
CD-15-00998	Deutschland	Human	NA	027	diese Arbeit	PRJEB33866
CD-15-00997	Deutschland	Human	NA	070	diese Arbeit	PRJEB33866
CD-15-00996	Deutschland	Human	NA	043	diese Arbeit	PRJEB33866
CD-15-00995	Deutschland	Human	NA	020	diese Arbeit	PRJEB33866
CD-15-00994	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-15-00993	Deutschland	Human	NA	014	diese Arbeit	PRJEB33866
CD-15-00992	Deutschland	Human	NA	003	diese Arbeit	PRJEB33866
CD-15-00991	Deutschland	Human	NA	043	diese Arbeit	PRJEB33866
CD-15-00990	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-15-00989	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-15-00988	Deutschland	Human	NA	015	diese Arbeit	PRJEB33866
CD-15-00987	Deutschland	Human	NA	126	diese Arbeit	PRJEB33866

CD-18-00566	Deutschland	Schwein	2012	n. d.	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00565	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00563	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780

CD-18-00562	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00561	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00560	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00559	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00548	Deutschland	Schwein	2012	049	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00543	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00542	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00541	Deutschland	Schwein	2012	049	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00538	Deutschland	Schwein	2012	049	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00537	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00536	Deutschland	Schwein	2012	049	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00533	Deutschland	Schwein	2012	598	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00532	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00530	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00528	Deutschland	Schwein	2012	598	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00527	Deutschland	Schwein	2012	598	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00525	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00524	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00520	Deutschland	Schwein	2012	598	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00519	Deutschland	Schwein	2012	598	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00518	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00517	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00516	Deutschland	Schwein	2012	598	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00515	Deutschland	Schwein	2012	598	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00514	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00513	Deutschland	Schwein	2012	598	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00510	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00509	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00508	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00506	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00503	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00502	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00501	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00500	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00499	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00498	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00497	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00496	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00495	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00494	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00493	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00492	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00490	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00489	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00488	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00487	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00486	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780

CD-18-00485	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00484	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00483	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00482	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00481	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00480	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00479	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00478	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00477	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00476	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00475	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00474	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00473	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00472	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00471	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00470	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00469	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00468	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00467	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00466	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00465	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00464	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00463	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00462	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00461	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00460	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00459	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00458	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00457	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00456	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00455	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00454	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00453	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00452	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00451	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00450	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00449	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00448	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00447	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00446	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00445	Deutschland	Schwein	2012	049	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00444	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00443	Deutschland	Schwein	2012	620	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00442	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00441	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00440	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00439	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780

CD-18-00438	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00437	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00436	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00435	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00434	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00433	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00432	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00431	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00430	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00429	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00428	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00427	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00426	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00425	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00424	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00423	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00422	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00421	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00420	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00419	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00418	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00417	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00377	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00376	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00375	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00374	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00373	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00371	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00370	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00369	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00368	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00367	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00366	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00365	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00364	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00363	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00362	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00361	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00360	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00359	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00358	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00357	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00356	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00355	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00354	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00353	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00352	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780

CD-18-00351	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00350	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00349	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00348	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00347	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00346	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00345	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00344	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00342	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00341	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00340	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00339	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00338	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00337	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00336	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00335	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00334	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00332	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00330	Deutschland	Schwein	2012	014	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00329	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00328	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00327	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00326	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00325	Deutschland	Schwein	2012	002	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00324	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00323	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00322	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00321	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00320	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00319	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00318	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00317	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00316	Deutschland	Schwein	2012	413	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00315	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00314	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00310	Deutschland	Schwein	2012	126	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00309	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00308	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00303	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780
CD-18-00302	Deutschland	Schwein	2012	078	Schneeberg et al. 2013[86]	PRJEB33780

Tabelle A.5: Liste der Isolate, die im Rahmend es SOARiAL Projektes isoliert wurden (Kapitel 2.2.1). Das Staubisolat wurde aus einer Probe des Windkanalversuchs isoliert, die anderen aus den entsprechenden Proben des Feldversuchs.

Isolatebezeichnung	Land	Isolationsquelle
CD-17-01062	Deutschland	Dünger
CD-17-01063	Deutschland	Dünger
CD-17-01064	Deutschland	Dünger
CD-17-01066	Deutschland	Dünger
CD-17-01067	Deutschland	Dünger
CD-17-01068	Deutschland	Dünger
CD-17-01069	Deutschland	Dünger
CD-17-01086	Deutschland	Sammelkotprobe
CD-17-01087	Deutschland	Sammelkotprobe
CD-17-01092	Deutschland	Sammelkotprobe
CD-17-01093	Deutschland	Dünger
CD-17-01094	Deutschland	Dünger
CD-17-01424	Deutschland	Dünger
CD-17-01475	Deutschland	Dünger
CD-17-01476	Deutschland	Dünger
CD-17-01477	Deutschland	Dünger
CD-17-01478	Deutschland	Dünger
CD-17-01479	Deutschland	Dünger
CD-17-01480	Deutschland	Dünger
CD-17-01482	Deutschland	Dünger
CD-17-01483	Deutschland	Dünger
CD-17-01484	Deutschland	Dünger
CD-17-01485	Deutschland	Dünger
CD-17-01486	Deutschland	Dünger
CD-17-01487	Deutschland	Dünger
CD-17-01488	Deutschland	Dünger
CD-17-01489	Deutschland	Dünger
CD-17-01491	Deutschland	Dünger
CD-17-01492	Deutschland	Dünger
CD-17-01493	Deutschland	Dünger
CD-17-01494	Deutschland	Dünger
CD-17-01495	Deutschland	Dünger
CD-17-01496	Deutschland	Dünger
CD-17-01497	Deutschland	Dünger
CD-17-01498	Deutschland	Dünger
CD-17-01499	Deutschland	Dünger
CD-17-01500	Deutschland	Dünger
CD-17-01501	Deutschland	Dünger
CD-17-01502	Deutschland	Dünger
CD-17-01503	Deutschland	Dünger
CD-17-01504	Deutschland	Dünger
CD-17-01505	Deutschland	Dünger
CD-17-01506	Deutschland	Dünger

CD-17-01510	Deutschland	Dünger
CD-17-01511	Deutschland	Dünger
CD-17-01512	Deutschland	Dünger
CD-17-01513	Deutschland	Dünger
CD-17-01514	Deutschland	Dünger
CD-17-01515	Deutschland	Dünger
CD-17-01516	Deutschland	Dünger
CD-17-01517	Deutschland	Dünger
CD-17-01518	Deutschland	Dünger
CD-17-01519	Deutschland	Dünger
CD-17-01520	Deutschland	Dünger
CD-17-01521	Deutschland	Dünger
CD-17-01522	Deutschland	Dünger
CD-17-01524	Deutschland	Dünger
CD-17-01525	Deutschland	Dünger
CD-17-01526	Deutschland	Dünger
CD-17-01527	Deutschland	Dünger
CD-17-01528	Deutschland	Dünger
CD-17-01529	Deutschland	Dünger
CD-17-01530	Deutschland	Dünger
CD-17-01532	Deutschland	Dünger
CD-17-01533	Deutschland	Staub
CD-17-01573	Deutschland	Dünger
CD-17-01574	Deutschland	Dünger
CD-17-01608	Deutschland	Dünger
CD-17-01609	Deutschland	Dünger
CD-17-01611	Deutschland	Dünger
CD-17-01612	Deutschland	Dünger
CD-17-01613	Deutschland	Dünger
CD-17-01614	Deutschland	Dünger
CD-17-01615	Deutschland	Dünger
CD-19-00275	Deutschland	Dünger
CD-19-00276	Deutschland	Dünger
CD-19-00277	Deutschland	gedüngter Boden
CD-19-00278	Deutschland	gedüngter Boden
CD-19-00279	Deutschland	gedüngter Boden
CD-19-00281	Deutschland	gedüngter Boden
CD-19-00282	Deutschland	gedüngter Boden
CD-19-00283	Deutschland	gedüngter Boden
CD-19-00284	Deutschland	gedüngter Boden
CD-19-00285	Deutschland	gedüngter Boden
CD-19-00286	Deutschland	gedüngter Boden
CD-19-00287	Deutschland	gedüngter Boden
CD-19-00288	Deutschland	gedüngter Boden
CD-19-00289	Deutschland	gedüngter Boden
CD-19-00290	Deutschland	gedüngter Boden
CD-19-00291	Deutschland	gedüngter Boden

CD-19-00292	Deutschland	gedüngter Boden
CD-19-00293	Deutschland	gedüngter Boden
CD-19-00294	Deutschland	gedüngter Boden
CD-19-00295	Deutschland	gedüngter Boden
CD-19-00296	Deutschland	gedüngter Boden
CD-19-00297	Deutschland	gedüngter Boden
CD-19-00298	Deutschland	gedüngter Boden
CD-19-00329	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00330	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00331	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00332	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00333	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00334	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00335	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00336	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00337	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00338	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00339	Deutschland	gedüngter Boden (2 Wochen)
CD-19-00340	Deutschland	gedüngter Boden (4 Wochen)
CD-19-00341	Deutschland	gedüngter Boden (4 Wochen)
CD-19-00342	Deutschland	gedüngter Boden (4 Wochen)
CD-19-00343	Deutschland	gedüngter Boden (4 Wochen)
CD-19-00344	Deutschland	gedüngter Boden (4 Wochen)
CD-19-00345	Deutschland	gedüngter Boden (4 Wochen)
CD-19-00346	Deutschland	gedüngter Boden (4 Wochen)
CD-19-00347	Deutschland	gedüngter Boden (4 Wochen)
CD-19-00352	Deutschland	gedüngter Boden (4 Wochen)
CD-19-00354	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00355	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00356	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00357	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00359	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00360	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00363	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00364	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00366	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00367	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00368	Deutschland	gedüngter Boden (7 Wochen)
CD-19-00373	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00374	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00375	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00376	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00378	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00379	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00380	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00382	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00383	Deutschland	gedüngter Boden (10 Wochen)

CD-19-00384	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00385	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00386	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00387	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00388	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00389	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00391	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00392	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00393	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00394	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00395	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00396	Deutschland	gedüngter Boden (10 Wochen)
CD-19-00397	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00398	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00399	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00400	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00401	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00402	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00403	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00404	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00406	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00408	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00409	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00410	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00411	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00412	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00413	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00414	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00416	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00417	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00418	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00419	Deutschland	gedüngter Boden (14 Wochen)
CD-19-00421	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00422	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00423	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00424	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00425	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00426	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00427	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00428	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00433	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00434	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00457	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00458	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00459	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00460	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00461	Deutschland	gedüngter Boden (19 Wochen)

CD-19-00462	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00463	Deutschland	gedüngter Boden (19 Wochen)
CD-19-00465	Deutschland	Boden
CD-19-00466	Deutschland	Boden
CD-19-00467	Deutschland	Boden
CD-19-00468	Deutschland	Boden
CD-19-00469	Deutschland	Boden

Anhang B

Höhere Populationsebenen von *C. difficile*

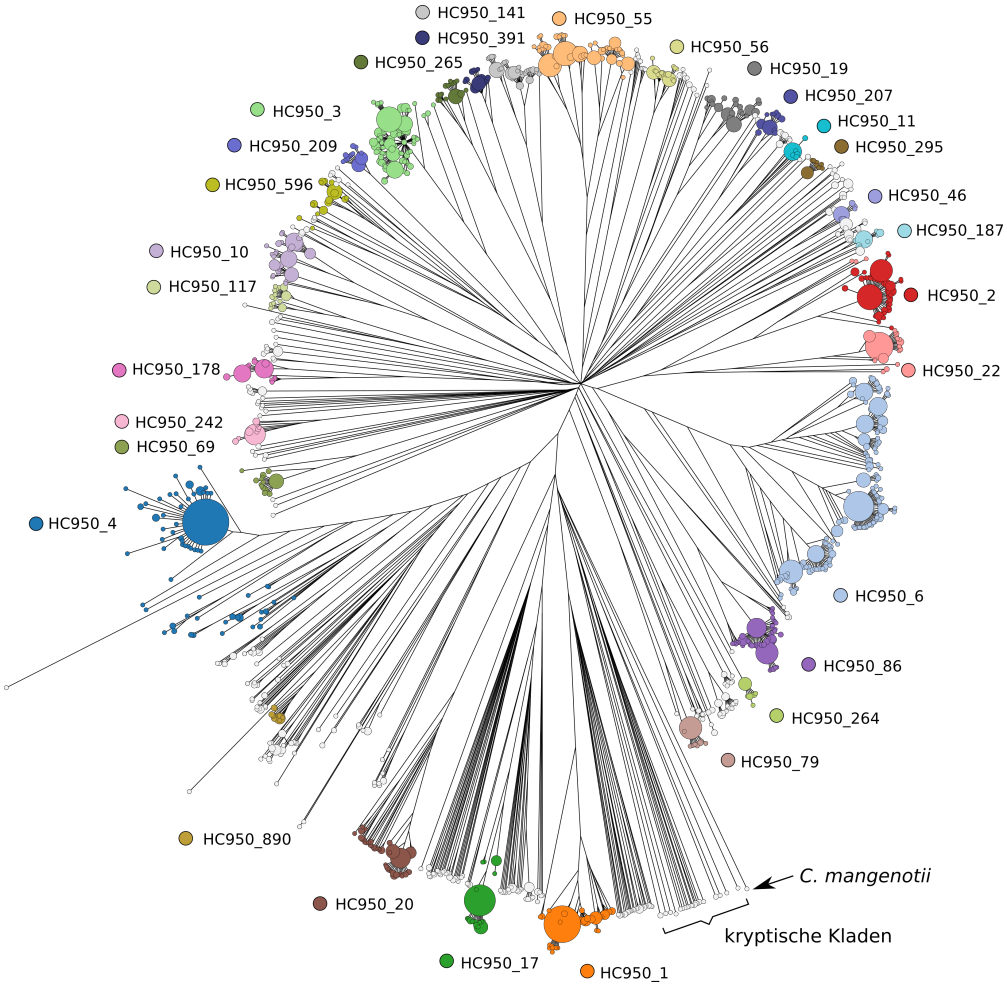

Abbildung B.1: Phylogenetischer Baum aller 13.515 *C. difficile* Genome in EnteroBase, eingefärbt nach HC950 Cluster. Der Baum basiert auf den cgMLST Allelprofilen der Genome und wurde mit dem Rapid-Neighbour-Joining Algorithmus berechnet.

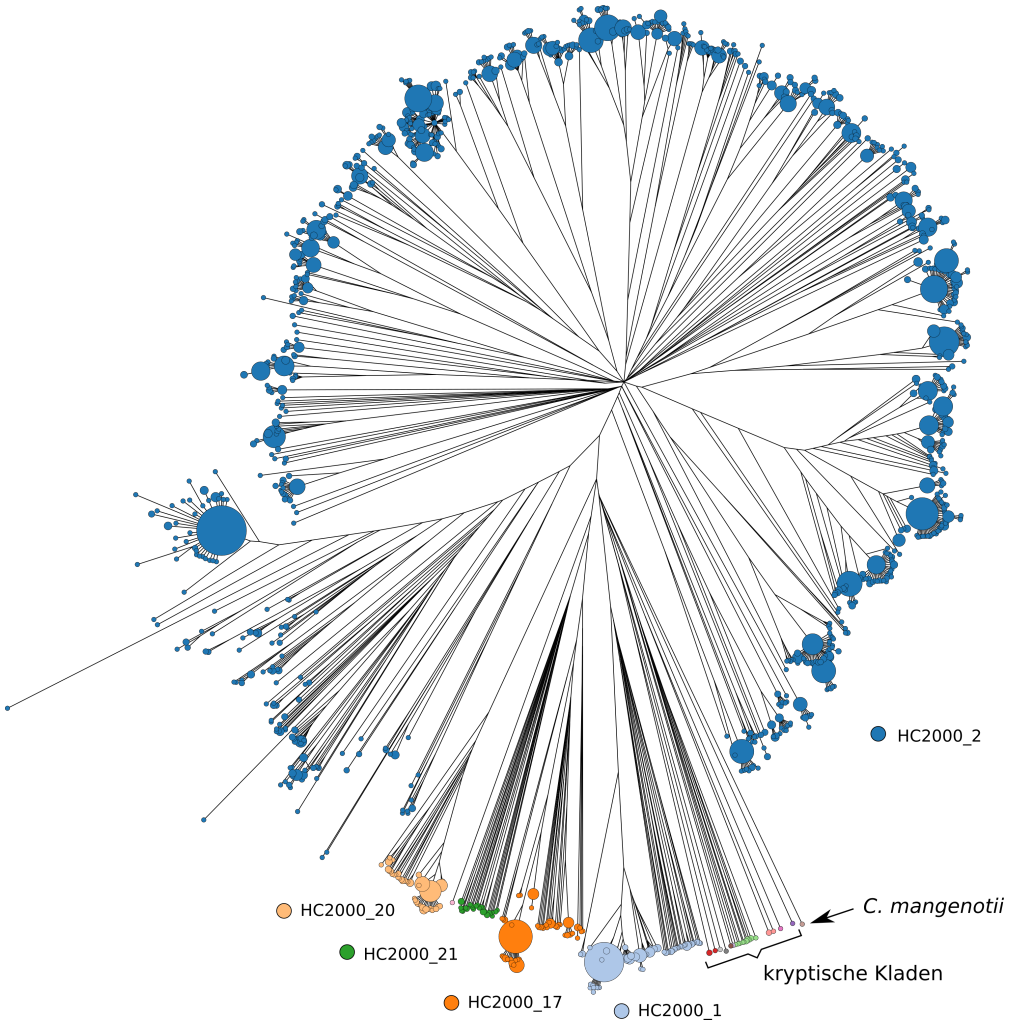

Abbildung B.2: Phylogenetischer Baum aller 13.515 *C. difficile* Genome in EnteroBase, eingefärbt nach HC2000 Cluster. Der Baum basiert auf den cgMLST Allelprofilen der Genome und wurde mit dem Rapid-Neighbour-Joining Algorithmus berechnet.

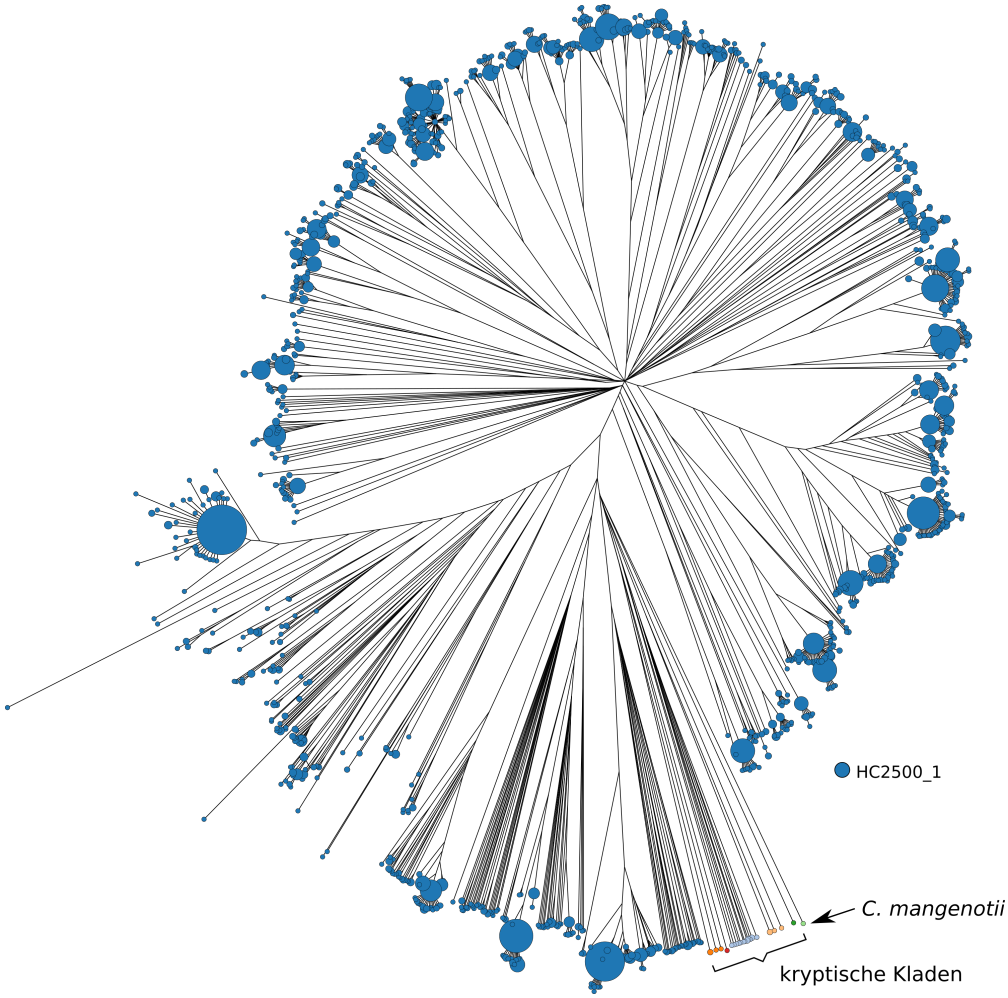

Abbildung B.3: Phylogenetischer Baum aller 13.515 *C. difficile* Genome in EnteroBase, eingefärbt nach HC2500 Cluster. Der Baum basiert auf den cgMLST Allelprofilen der Genome und wurde mit dem Rapid-Neighbour-Joining Algorithmus berechnet.

Anhang C

Ergänzende Tabellen

Tabelle C.1: SNP Distanzen basierend auf *Read-Mappings* gegen verschiedene Referenzsequenzen für paarweise Vergleiche zwischen Genomen der Krankenhausisolate aus Kapitel 3.4.2 und einem anderen Genom in EnteroBase. Diese Beziehungen konnten in insgesamt 15 HC2 Cluster gefunden werden.

klinisches Isolat aus Kapitel 3.4.2	Eintrag in EnteroBase	SNP Distanzen				Jahres-differenz	HC2 Cluster	wgMLST Allel-unterschiede
		R20291 (CC4)	CD196 (CC4)	CD630 (CC58)	M120 (CC1)			
CD-16-00166	CLO_AA8571AA	3	3	1	9	3	1 (CC1)	0,0648
CD-16-00166	CLO_AA3704AA	2	2	2	8	4	1 (CC1)	0,0852
CD-16-00166	CLO_AA3710AA	4	4	5	11	4	1 (CC1)	0,0724
CD-16-00173	CLO_AA3751AA	4	4	3	5	4	1 (CC1)	0,1804
CD-15-00941	CLO_AA9728AA	2	2	2	0	1	76 (CC3)	0,1318
CD-15-00941	CLO_AA9603AA	3	3	3	0	1	76 (CC3)	0,023
CD-15-00941	CLO_BA0019AA	0	0	0	0	1	76 (CC3)	0,1343
CD-15-00941	CLO_AA9634AA	2	2	2	1	1	76 (CC3)	0,043
CD-15-00941	CLO_AA9808AA	0	0	0	0	1	76 (CC3)	0,1351
CD-15-00941	CLO_AA9698AA	1	1	2	1	1	76 (CC3)	0,1248
CD-15-00639	CLO_AA4199AA	0	0	0	0	1	109 (CC3)	0
CD-15-00727	CLO_AA4557AA	8	9	10	3	NA	109 (CC3)	0,2961
CD-15-00727	CLO_AA4558AA	8	9	10	3	NA	109 (CC3)	0,2966
CD-15-00727	CLO_AA4303AA	19	20	18	5	NA	109 (CC3)	0,2269
CD-15-00727	CLO_AA4304AA	19	20	18	5	NA	109 (CC3)	0,2269
CD-15-00727	CLO_AA7713AA	8	9	9	3	5	109 (CC3)	0,2492
CD-15-00727	CLO_BA6868AA	10	11	11	3	1	109 (CC3)	0,1705
CD-15-00727	CLO_BA6410AA	10	11	10	4	5	109 (CC3)	0,1848
CD-15-00727	CLO_BA6462AA	13	14	13	3	6	109 (CC3)	0,2593
CD-15-00727	CLO_BA4170AA	9	10	10	3	6	109 (CC3)	0,3052
CD-15-00734	CLO_AA4557AA	8	9	10	3	NA	109 (CC3)	0,3013
CD-15-00734	CLO_AA4558AA	8	9	10	3	NA	109 (CC3)	0,302
CD-15-00734	CLO_BA6868AA	10	11	11	3	1	109 (CC3)	0,1846
CD-15-00734	CLO_BA4170AA	9	10	10	3	6	109 (CC3)	0,259
CD-15-00727	CLO_AA9436AA	17	18	16	5	1	109 (CC3)	0,2339
CD-15-00727	CLO_AA2319AA	8	9	8	3	NA	109 (CC3)	0,2007
CD-15-00727	CLO_AA4721AA	8	9	8	3	7	109 (CC3)	0,1942
CD-15-00727	CLO_AA4690AA	11	12	11	3	5	109 (CC3)	0,2196

CD-15-00727	CLO_AA2747AA	8	9	8	3	NA	109 (CC3)	0,194
CD-15-00727	CLO_AA4722AA	8	9	8	3	7	109 (CC3)	0,1957
CD-15-00727	CLO_AA2751AA	8	9	8	3	NA	109 (CC3)	0,192
CD-15-00727	CLO_AA2865AA	8	9	8	3	NA	109 (CC3)	0,1922
CD-15-00727	CLO_AA2791AA	8	9	8	3	NA	109 (CC3)	0,1808
CD-15-00961	CLO_BA6818AA	9	9	12	5	7	282 (CC6)	0,0938
CD-16-00088	CLO_AA6002AA	14	14	9	4	6	310 (CC4)	0
CD-16-00088	CLO_BA6554AA	11	11	8	4	9	310 (CC4)	0,0047
CD-16-00088	CLO_BA6300AA	12	12	9	4	9	310 (CC4)	0,0047
CD-16-00088	CLO_BA4965AA	6	6	6	2	8	310 (CC4)	0,016
CD-16-00088	CLO_AA0574AA	3	3	3	2	8	310 (CC4)	0,0009
CD-16-00088	CLO_AA0573AA	5	5	5	3	8	310 (CC4)	0,0056
CD-16-00088	CLO_AA0571AA	3	3	3	2	7	310 (CC4)	0
CD-16-00088	CLO_AA0570AA	3	3	3	2	7	310 (CC4)	0,0009
CD-16-00088	CLO_AA0568AA	3	3	3	2	7	310 (CC4)	0,0038
CD-16-00088	CLO_AA0598AA	2	2	2	1	7	310 (CC4)	0,0047
CD-16-00088	CLO_AA0564AA	2	2	2	1	6	310 (CC4)	0,0047
CD-16-00088	CLO_AA0596AA	2	2	2	1	6	310 (CC4)	0
CD-16-00088	CLO_AA0561AA	5	4	4	2	6	310 (CC4)	0
CD-16-00088	CLO_AA0594AA	2	2	2	1	5	310 (CC4)	0,0047
CD-16-00088	CLO_AA0558AA	2	2	2	1	5	310 (CC4)	0,0047
CD-16-00088	CLO_AA0556AA	7	7	6	2	5	310 (CC4)	0,0056
CD-16-00088	CLO_AA0543AA	6	5	5	2	7	310 (CC4)	0
CD-16-00088	CLO_AA0554AA	6	5	4	2	6	310 (CC4)	0,0047
CD-16-00088	CLO_AA0583AA	7	7	4	3	3	310 (CC4)	0,0047
CD-16-00088	CLO_AA0544AA	5	4	4	2	3	310 (CC4)	0,0112
CD-16-00088	CLO_AA0582AA	4	4	4	2	3	310 (CC4)	0,0093
CD-16-00088	CLO_AA0581AA	5	4	4	2	3	310 (CC4)	0,0103
CD-16-00088	CLO_AA3573AA	11	10	6	2	NA	310 (CC4)	0,0065
CD-15-00856	CLO_AA7218AA	4	4	4	3	1	479 (CC2)	0,0618
CD-15-00856	CLO_AA7215AA	4	4	4	3	1	479 (CC2)	0,0618
CD-15-00856	CLO_BA6787AA	7	7	5	4	7	479 (CC2)	0,1119
CD-15-00859	CLO_AA7218AA	4	4	4	3	1	479 (CC2)	0,0608
CD-15-00859	CLO_AA7215AA	4	4	4	3	1	479 (CC2)	0,067
CD-15-00859	CLO_BA6787AA	7	7	5	4	7	479 (CC2)	0,118
CD-15-00856	CLO_AA0865AA	5	5	3	2	3	479 (CC2)	0,1101
CD-15-00856	CLO_AA3529AA	6	6	4	3	NA	479 (CC2)	0,1411
CD-15-00859	CLO_AA0865AA	5	5	3	2	3	479 (CC2)	0,117
CD-15-00859	CLO_AA3529AA	6	6	4	3	NA	479 (CC2)	0,1429
CD-15-00638	CLO_AA9916AA	6	6	4	2	0	1131 (CC4)	0,0129
CD-15-00638	CLO_AA9555AA	1	1	1	1	0	1131 (CC4)	0,0166
CD-15-00638	CLO_AA9637AA	5	5	5	4	0	1131 (CC4)	0,0379
CD-15-00638	CLO_AA9829AA	2	2	3	0	0	1131 (CC4)	0,0037
CD-15-00638	CLO_AA9935AA	2	2	2	2	0	1131 (CC4)	0,0531
CD-15-00857	CLO_AA0545AA	0	0	0	0	2	1131 (CC4)	0,0243
CD-16-00168	CLO_AA4539AA	3	3	2	5	NA	1178 (CC114)	0
CD-15-00966	CLO_AA4313AA	3	3	3	1	NA	1195 (CC71)	0,0569

CD-15-00966	CLO_AA0501AA	4	3	4	2	NA	1195 (CC71)	0,0413
CD-15-00966	CLO_AA0499AA	4	3	4	2	NA	1195 (CC71)	0,0421
CD-15-00966	CLO_AA0498AA	4	3	4	2	NA	1195 (CC71)	0,0421
CD-15-00733	CLO_AA9511AA	3	3	2	1	1	1232 (CC3)	0,0058
CD-16-00161	CLO_AA9511AA	3	3	2	1	2	1232 (CC3)	0,0058
CD-15-00673	CLO_AA1780AA	3	3	1	2	4	1243 (CC264)	0,006
CD-15-00673	CLO_AA1743AA	3	3	1	2	NA	1243 (CC264)	0,0076
CD-15-00736	CLO_AA1780AA	3	3	1	2	5	1243 (CC264)	0,006
CD-15-00736	CLO_AA1743AA	3	3	1	2	NA	1243 (CC264)	0,0051
CD-15-00674	CLO_AA9749AA	2	2	2	1	0	1276 (CC71)	0,0075
CD-15-00674	CLO_BA0016AA	0	0	0	0	0	1276 (CC71)	0,0076
CD-16-00142	CLO_BA0858AA	6	6	12	1	2	2418 (CC10)	0,0357
CD-16-00142	CLO_BA0256AA	9	9	10	5	2	2418 (CC10)	0,005
CD-16-00142	CLO_BA0951AA	6	6	12	1	2	2418 (CC10)	0,0359
CD-15-00685	CLO_AA3952AA	2	2	3	5	7	3262 (CC1)	0,1325
CD-16-00180	CLO_AA7714AA	4	4	2	2	6	4456 (CC3)	0

Tabelle C.2: Kontingenztafel für die Genome der Isolate, die in einem regionalen Netzwerk von Krankenhäusern isoliert wurden. 133 Isolate dieser Studie wurden durch die cgMLST-Analyse in 23 HC2 Cluster eingeteilt. Die Tabelle zeigt die Verteilung der Isolate pro HC2 Cluster über die Stationen der beprobten Krankenhäuser an.

Station / HC2 Cluster	1_1	1_2	1_3	1_4	2_1	2_11	2_12	2_2	2_3	2_5	2_6	2_8	2_9	3_1	3_10	3_11	3_12	3_13	3_14	3_2
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
76	7	4	1	3	0	0	0	1	0	9	0	0	3	0	1	1	1	1	1	4
85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
109	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
479	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
491	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1127	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1131	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1206	0	1	0	0	0	2	1	0	0	0	0	0	0	0	1	0	0	0	0	0
1208	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1210	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
1225	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1232	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1242	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1243	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1251	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
1267	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
4415	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4431	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4808	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4823	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0

Station / HC2 / Cluster	3_3	3_4	3_5	3_6	3_7	3_8	3_9	4_1	4_2	4_3	4_4	5_10	5_11	5_13	5_14	5_15	5_17	5_18	5_19	5_2
1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
76	1	1	0	4	1	0	3	0	0	1	1	0	1	0	0	1	1	0	0	0
85	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
109	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
479	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
491	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1127	0	0	0	0	0	0	0	1	0	0	0	0	0	3	0	0	0	0	0	0
1131	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0
1206	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1208	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1210	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1225	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1232	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1242	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1243	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1251	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1267	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4415	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4431	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
4808	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4823	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Station / HC2 Cluster	5_23	5_26	5_27	5_28	5_29	5_31	5_5	5_6	5_7	5_8	6_1	6_2	6_3	6_4	NA
1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
76	0	1	1	1	0	1	1	0	0	0	4	2	1	1	1
85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
109	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0
479	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
491	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1127	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0
1131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1206	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
1208	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1210	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1225	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1232	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
1242	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1243	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
1251	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
1267	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4415	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4431	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4808	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
4823	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Literatur

[1] Ö. Açikgöz und A. Günay, "The early impact of the Covid-19 pandemic on the global and Turkish economy", *Turkish Journal of Medical Sciences*, Jg. 50, Nr. SI-1, S. 520–526, 2020. DOI: 10.3906/sag-2004-6.

[2] A. Wilder-Smith und D. O. Freedman, "Isolation, quarantine, social distancing and community containment: Pivotal role for old-style public health measures in the novel coronavirus (2019-nCoV) outbreak", *Journal of Travel Medicine*, Jg. 27, Nr. 2, März 2020. DOI: 10.1093/jtm/taaa020.

[3] C. Watson, "How countries are using genomics to help avoid a second coronavirus wave", *Nature*, Jg. 582, Nr. 7810, S. 19, Juni 2020. DOI: 10.1038/d41586-020-01573-5.

[4] Y. Z. Zhang und E. C. Holmes, "A genomic perspective on the origin and emergence of SARS-CoV-2", *Cell*, Jg. 181, Nr. 2, S. 223–227, Apr. 2020. DOI: 10.1016/j.cell.2020.03.035.

[5] S. Reardon, "Antibiotic treatment for COVID-19 complications could fuel resistant bacteria", *Science*, Apr. 2020. DOI: 10.1126/science.abc2995.

[6] W.-j. Guan, Z.-y. Ni, Y. Hu, W. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui, B. Du, L. Li, G. Zeng, K. Y. Yuen, R. Chen, C.-l. Tang, T. Wang, P. Chen, J. Xiang, S. Li, J. L. Wang, Z.-j. Liang, Y. Peng, L. Wei, Y. Liu, Y. H. Hu, P. Peng, J. M. Wang, J.-y. Liu, Z. Chen, G. Li, Z.-j. Zheng, S.-q. Qiu, J. Luo, C.-j. Ye, S.-y. Zhu und N.-s. Zhong, "Clinical characteristics of coronavirus disease 2019 in China", *New England Journal of Medicine*, Jg. 382, Nr. 18, S. 1708–1720, Apr. 2020. DOI: 10.1056/NEJMoa2002032.

[7] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, L. Guan, Y. Wei, H. Li, X. Wu, J. Xu, S. Tu, Y. Zhang, H. Chen und B. Cao, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study", *The Lancet*, Jg. 395, Nr. 10229, S. 1054–1062, März 2020. DOI: 10.1016/S0140-6736(20)30566-3.

[8] A. Sandhu, G. Tillotson, J. Polistico, H. Salimnia, M. Cranis, J. Moshos, L. Cullen, L. Jabbo, L. Diebel und T. Chopra, "*Clostridioides difficile* in COVID-19 patients, Detroit, Michigan, USA, March–April 2020", *Emerging Infectious Diseases*, Jg. 26, Nr. 9, Sep. 2020. DOI: 10.3201/eid2609.202126.

[9] CDC, "Antibiotic resistance threats in the United States, 2019", *U.S. Department of Health and Human Services, CDC*, 2019. Adresse: http://dx.doi.org/10.15620/cdc:82532.

[10] R. Capita und C. Alonso-Calleja, "Antibiotic-resistant bacteria: A challenge for the food industry", *Critical Reviews in Food Science and Nutrition*, Jg. 53, Nr. 1, S. 11–48, Jan. 2013. DOI: 10.1080/10408398.2010.519837.

[11] M. H. Kollef und V. J. Fraser, "Antibiotic resistance in the intensive care unit", *Annals of Internal Medicine*, Jg. 134, Nr. 4, S. 298–314, Feb. 2001. DOI: 10.7326/0003-4819-134-4-200102200-00014.

[12] I. N. Okeke und R. Edelman, "Dissemination of antibiotic-resistant bacteria across geographic borders", *Clinical Infectious Diseases*, Jg. 33, Nr. 3, S. 364–369, Aug. 2001. DOI: 10.1086/321877.

[13] K. A. Brown, N. Khanafer, N. Daneman und D. N. Fisman, "Meta-analysis of antibiotics and the risk of community-associated *Clostridium difficile* infection", *Antimicrobial Agents and Chemotherapy*, Jg. 57, Nr. 5, S. 2326–2332, Mai 2013. DOI: 10.1128/AAC.02176-12.

[14] R. Koch-Institut, "Hygienemaßnahmen bei *Clostridioides difficile*-Infektion (CDI): Empfehlung der Kommission für Krankenhaushygiene und Infektionsprävention (KRINKO) beim Robert Koch-Institut", *Bundesgesundheitsblatt*, Jg. 62, Nr. 7, S. 906–923, 2019. DOI: 10.1007/s00103-019-02959-1.

[15] M. He, F. Miyajima, P. Roberts, L. Ellison, D. J. Pickard, M. J. Martin, T. R. Connor, S. R. Harris, D. Fairley, K. B. Bamford, S. D'Arc, J. Brazier, D. Brown, J. E. Coia, G. Douce, D. Gerding, H. J. Kim, T. H. Koh, H. Kato, M. Senoh, T. Louie, S. Michell, E. Butt, S. J. Peacock, N. M. Brown, T. Riley, G. Songer, M. Wilcox, M. Pirmohamed, E. Kuijper, P. Hawkey, B. W. Wren, G. Dougan, J. Parkhill und T. D. Lawley, "Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*.", *Nature genetics*, Jg. 45, Nr. 1, S. 109–13, 2013. DOI: 10.1038/ng.2478.

[16] M. Steglich, A. Nitsche, L. Von Müller, M. Herrmann, T. A. Kohl, S. Niemann und U. Nübel, "Tracing the spread of *Clostridium difficile* ribotype 027 in Germany based on bacterial genome sequences", *PLoS ONE*, Jg. 10, Nr. 10, S. 1–11, 2015. DOI: 10.1371/journal.pone.0139811.

[17] I. C. Hall und E. O'Toole, "Intestinal flora in newborn infants with a description of a new pathogenic anaerobe *Bacillus difficilis*", *Am J Dis Child*, Jg. 49, S. 390, 1935.

[18] M. Y. Galperin, V. Brover, I. Tolstoy und N. Yutin, "Phylogenomic analysis of the family *Peptostreptococcaceae* (Clostridium cluster XI) and proposal for reclassification of *Clostridium litorale* (Fendrich et al. 1991) and *Eubacterium acidaminophilum* (Zindel et al. 1989) as *Peptoclostridium litorale* gen. nov. comb. nov. and *Peptoclostridium acidaminophilum* comb. nov.", *International Journal of Systematic and Evolutionary Microbiology*, Jg. 66, Nr. 12, S. 5506–5513, Dez. 2016. DOI: 10.1099/ijsem.0.001548.

[19] P. A. Lawson, D. M. Citron, K. L. Tyrrell und S. M. Finegold, "Reclassification of *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1935) Prévot 1938", *Anaerobe*, Jg. 40, S. 95–99, Aug. 2016. DOI: 10.1016/j.anaerobe.2016.06.008.

[20] F. J. Tedesco, R. W. Barton und D. H. Alpers, "Clindamycin-associated colitis - A prospective study", *Annals of Internal Medicine*, Jg. 81, Nr. 4, S. 429–433, 1974. DOI: 10.7326/0003-4819-81-4-429.

[21] J. G. Bartlett, t. W. Chang, M. Gurwith, S. L. Gorbach und A. B. Onderdonk, "Antibiotic-associated pseudomembranous colitis due to toxin-producing *Clostridia*", *New England Journal of Medicine*, Jg. 298, Nr. 10, S. 531–534, März 1978. DOI: 10.1056/NEJM197803092981003.

[22] D. R. Knight, B. Elliott, B. J. Chang, T. T. Perkins und T. V. Riley, "Diversity and evolution in the genome of *Clostridium difficile*", *Clinical Microbiology Reviews*, Jg. 28, Nr. 3, S. 721–741, 2015. DOI: 10.1128/CMR.00127-14.

[23] J. G. Bartlett, "Antibiotic-associated diarrhea", *New England Journal of Medicine*, Jg. 346, Nr. 5, S. 334–339, Jan. 2002. DOI: 10.1056/NEJMcp011603.

[24] I. Adlerberth, H. Huang, E. Lindberg, N. Åberg, B. Hesselmar, R. Saalman, C. E. Nord, A. E. Wold und A. Weintraubb, "Toxin-producing *Clostridium difficile* strains as long-term gut colonizers in healthy infants", *Journal of Clinical Microbiology*, Jg. 52, Nr. 1, S. 173–179, Jan. 2014. DOI: 10.1128/JCM.01701-13.

[25] T. Blixt, K. O. Gradel, C. Homann, J. B. Seidelin, K. Schønning, A. Lester, J. Houlind, M. Stangerup, M. Gottlieb und J. D. Knudsen, "Asymptomatic carriers contribute to nosocomial *Clostridium difficile* infection: A cohort study of 4508 patients", *Gastroenterology*, Jg. 152, Nr. 5, 1031–1041.e2, Apr. 2017. DOI: 10.1053/j.gastro.2016.12.035.

[26] S. S. Magill, J. R. Edwards, W. Bamberg, Z. G. Beldavs, G. Dumyati, M. A. Kainer, R. Lynfield, M. Maloney, L. McAllister-Hollod, J. Nadle, S. M. Ray, D. L. Thompson, L. E. Wilson und S. K. Fridkin, "Multistate point-prevalence survey of health care-associated infections", *New England Journal of Medicine*, Jg. 370, Nr. 13, S. 1198–1208, 2014. DOI: 10.1056/NEJMoa1306801.

[27] B. Elliott, G. O. Androga, D. R. Knight und T. V. Riley, "*Clostridium difficile* infection: Evolution, phylogeny and molecular epidemiology", *Infection, Genetics and Evolution*, Jg. 49, S. 1–11, 2017. DOI: 10.1016/j.meegid.2016.12.018.

[28] R. Rodrigues, G. E. Barber und A. N. Ananthakrishnan, "A comprehensive study of costs associated with recurrent *Clostridium difficile* infection", *Infection Control and Hospital Epidemiology*, Jg. 38, Nr. 2, S. 196–202, Feb. 2017. DOI: 10.1017/ice.2016.246.

[29] S. Janezic, M. Potocnik, V. Zidaric und M. Rupnik, "Highly divergent *Clostridium difficile* strains isolated from the environment", *PLoS ONE*, Jg. 11, Nr. 11, D. Paredes-Sabja, Hrsg., e0167101, Nov. 2016. DOI: 10.1371/journal.pone.0167101.

[30] B. B. Lewis, R. A. Carter, L. Ling, I. Leiner, Y. Taur, M. Kamboj, E. R. Dubberke, J. Xavier und E. G. Pamer, "Pathogenicity locus, core genome, and accessory gene contributions to *Clostridium difficile* virulence", *mBio*, Jg. 8, Nr. 4, e00885–17, Sep. 2017. DOI: 10.1128/mBio.00885-17.

[31] J. Scaria, L. Ponnala, T. Janvilisri, W. Yan, L. A. Mueller und Y. F. Chang, "Analysis of ultra low genome conservation in *Clostridium difficile*", *PLoS ONE*, Jg. 5, Nr. 12, M. J. Horsburgh, Hrsg., e15147, Dez. 2010. DOI: 10.1371/journal.pone.0015147.

[32] C. Lübbert, E. John und L. von Müller, "*Clostridium difficile* infection: Guideline-based diagnosis and treatment", *Deutsches Ärzteblatt international*, Jg. 111, Nr. 43, S. 723–731, Okt. 2014. DOI: 10.3238/arztebl.2014.0723.

[33] K. Sánchez-Hurtado, M. Corretge, E. Mutlu, R. McIlhagger, J. M. Starr und I. R. Poxton, "Systemic antibody response to *Clostridium difficile* in colonized patients with and without symptoms and matched controls", *Journal of Medical Microbiology*, Jg. 57, Nr. 6, S. 717–724, Juni 2008. DOI: 10.1099/jmm.0.47713-0.

[34] G. E. Bignardi, "Risk factors for *Clostridium difficile* infection", *Journal of Hospital Infection*, Jg. 40, Nr. 1, S. 1–15, 1998. DOI: 10.1016/S0195-6701(98)90019-6.

[35] M. Monot, C. Eckert, A. Lemire, A. Hamiot, T. Dubois, C. Tessier, B. Dumoulard, B. Hamel, A. Petit, V. Lalande, L. Ma, C. Bouchier, F. Barbut und B. Dupuy, "*Clostridium difficile*: New insights into the evolution of the pathogenicity locus", *Scientific reports*, Jg. 5, S. 15 023, Okt. 2015. DOI: 10.1038/srep15023.

[36] D. N. Gerding, S. Johnson, M. Rupnik und K. Aktories, "*Clostridium difficile* binary toxin CDT: Mechanism, epidemiology, and potential clinical importance", *Gut microbes*, Jg. 5, Nr. 1, S. 15–27, 2014. DOI: 10.4161/gmic.26854.

[37] P. Moore, L. Kyne, A. Martin und K. Solomon, "Germination efficiency of clinical *Clostridium difficile* spores and correlation with ribotype, disease severity and therapy failure", *Journal of Medical Microbiology*, Jg. 62, Nr. Pt 9, S. 1405–1413, 2013. DOI: 10.1099/jmm.0.056614-0.

[38] L. K. Kociolek und D. N. Gerding, "Breakthroughs in the treatment and prevention of *Clostridium difficile* infection", *Nature Reviews Gastroenterology and Hepatology*, Jg. 13, Nr. 3, S. 150–160, März 2016. DOI: 10.1038/nrgastro.2015.220.

[39] I. Figueroa, S. Johnson, S. P. Sambol, E. J. Goldstein, D. M. Citron und D. N. Gerding, "Relapse versus reinfection: Recurrent *Clostridium difficile* infection following treatment with fidaxomicin or vancomycin", *Clinical Infectious Diseases*, Jg. 55, Nr. Suppl 2, S104–9, Aug. 2012. DOI: 10.1093/cid/cis357.

[40] D. W. Eyre, F. Babakhani, D. Griffiths, J. Seddon, C. Del Ojo Elias, S. L. Gorbach, T. E. Peto, D. W. Crook und A. S. Walker, "Whole-genome sequencing demonstrates that fidaxomicin is superior to vancomycin for preventing reinfection and relapse of infection with *Clostridium difficile*", *Journal of Infectious Diseases*, Jg. 209, Nr. 9, S. 1446–1451, 2014. DOI: 10.1093/infdis/jit598.

[41] A. Durovic, A. F. Widmer, R. Frei und S. Tschudin-Sutter, "Distinguishing *Clostridium difficile* recurrence from reinfection: Independent validation of current recommendations", *Infection Control and Hospital Epidemiology*, Jg. 38, Nr. 8, S. 891–896, Aug. 2017. DOI: 10.1017/ice.2017.119.

[42] T. A. Rubin, C. E. Gessert, J. Aas und J. S. Bakken, "Fecal microbiome transplantation for recurrent *Clostridium difficile* infection: Report on a case series", *Anaerobe*, Jg. 19, Nr. 1, S. 22–26, Feb. 2013. DOI: 10.1016/j.anaerobe.2012.11.004.

[43] R. Koch-institut, "Surveillance von nosokomialen Infektionen : Empfehlung der Kommission für Krankenhaushygiene und Infektionsprävention (KRINKO) beim Robert Koch-Institut", *Bundesgesundheitsblatt*, Jg. 63, Nr. 2, S. 228–241, 2020. DOI: 10.1007/s00103-019-03077-8.

[44] B. Blümel, D. Sagebiel, A. Gilsdorf und M. Diercke, "Positive predictive value of the German notification system for infectious diseases: Surveillance data from eight local health departments, Berlin, 2012", *PLoS ONE*, Jg. 14, Nr. 2, 2019. DOI: 10.1371/journal.pone.0212908.

[45] A. van Belkum, P. T. Tassios, L. Dijkshoorn, S. Haeggman, B. Cookson, N. K. Fry, V. Fussing, J. Green, E. Feil, P. Gerner-smidt, S. Brisse und M. Struelens, "Guidelines for the validation and application of typing methods for use in bacterial epidemiology", *Clinical Microbiology and Infection*, Jg. 13, Nr. Suppl. 3, S. 1–46, 2007. DOI: 10.1111/j.1469-0691.2007.01786.x.

[46] S. J. Peacock, J. Parkhill und N. M. Brown, "Changing the paradigm for hospital outbreak detection by leading with genomic surveillance of nosocomial pathogens", *Microbiology*, Jg. 164, Nr. 10, S. 1213–1219, 2018. DOI: 10.1099/mic.0.000700.

[47] M. K. Hayden, "Detection of nosocomial outbreaks: genomic surveillance takes the lead", *Clinical Infectious Diseases*, Jg. 3, S. 3–5, 2019. DOI: 10.1093/cid/ciz667.

[48] C. W. Knetsch, T. D. Lawley, M. P. Hensgens, J. Corver, M. W. Wilcox und E. J. Kuijper, "Current application and future perspectives of molecular typing methods to study *Clostridium difficile* infections", *Eurosurveillance*, Jg. 18, Nr. 4, 2013. DOI: 10.2807/ese.18.04.20381-en.

[49] S. H. Cohen, Y. J. Tang und J. Silva, "Molecular typing methods for the epidemiological identification of *Clostridium difficile* strains", *Expert Review of Molecular Diagnostics*, Jg. 1, Nr. 1, S. 61–70, 2001, ISSN: 14737140. DOI: 10.1586/14737159.1.1.61.

[50] S. Janezic und M. Rupnik, "Molecular typing methods for *Clostridium difficile*: Pulsed-field gel electrophoresis and PCR ribotyping", *Methods in Molecular Biology*, Jg. 646, S. 55–65, 2010. DOI: 10.1007/978-1-60327-365-7_4.

[51] L. Uelze, J. Grützke, M. Borowiak, J. A. Hammerl, K. Juraschek, C. Deneke, S. H. Tausch und B. Malorny, "Typing methods based on whole genome sequencing data", *One Health Outlook*, Jg. 2, Nr. 1, S. 1–19, Dez. 2020. DOI: 10.1186/s42522-020-0010-1.

[52] W. Kallow, M. Erhard, H. N. Shah, E. Raptakis und M. Welker, "MALDI-TOF MS for microbial identification: Years of experimental development to an established protocol", *Mass Spectrometry for Microbial Proteomics*, S. 255–276, Juni 2010. DOI: 10.1002/9780470665497.ch12.

[53] M. F. Emele, F. M. Joppe, T. Riedel, J. Overmann, M. Rupnik, P. Cooper, R. L. Kusumawati, F. K. Berger, F. Laukien, O. Zimmermann, W. Bohne, U. Groß, O. Bader und A. E. Zautner, "Proteotyping of *Clostridioides difficile* as alternate typing method to ribotyping is able to distinguish the ribotypes RT027 and RT176 from other ribotypes", *Frontiers in Microbiology*, Jg. 10, S. 2087, Sep. 2019. DOI: 10.3389/fmicb.2019.02087.

[54] A. Indra, S. Huhulescu, M. Schneeweis, P. Hasenberger, S. Kernbichler, A. Fiedler, G. Wewalka, F. Allerberger und E. J. Kuijper, "Characterization of *Clostridium difficile* isolates using capillary gel electrophoresis-based PCR ribotyping", *Journal of medical microbiology*, Jg. 57, Nr. 11, S. 1377–82, Nov. 2008. DOI: 10.1099/jmm.0.47714-0.

[55] S. García-Fernández, M. Frentrup, M. Steglich, A. Gonzaga, M. Cobo, N. López-Fresneña, J. Cobo, M. I. Morosini, R. Cantón, R. del Campo und U. Nübel, "Whole-genome sequencing reveals nosocomial *Clostridioides difficile* transmission and a previously unsuspected epidemic scenario", *Scientific Reports*, Jg. 9, Nr. 1, S. 1–9, 2019. DOI: 10.1038/s41598-019-43464-4.

[56] D. R. Knight, B. Kullin, G. O. Androga, F. Barbut, C. Eckert, S. Johnson, P. Spigaglia, K. Tateda, P.-J. Tsai und T. V. Riley, "Evolutionary and genomic insights into *Clostridioides difficile* sequence type 11: a diverse zoonotic and antimicrobial-resistant lineage of global One Health importance", *mBio*, Jg. 10, Nr. 2, 2019. DOI: 10.1128/mbio.00446-19.

[57] N. H. Zaiß, "Molekulare Epidemiologie und Populationsbiologie von *Clostridium difficile*", *Dissertation, Technische Universität Carolo-Wilhelmina zu Braunschweig*, 2010.

[58] R. T. Espejo und N. Plaza, "Multiple ribosomal RNA operons in bacteria; their concerted evolution and potential consequences on the rate of evolution of their 16S rRNA", *Frontiers in Microbiology*, Jg. 9, Nr. JUN, Juni 2018. DOI: 10.3389/fmicb.2018.01232.

[59] W. N. Fawley, C. W. Knetsch, D. R. MacCannell, C. Harmanus, T. Du, M. R. Mulvey, A. Paulick, L. Anderson, E. J. Kuijper und M. H. Wilcox, "Development and validation of an internationally-standardized, high-resolution capillary gel-based electrophoresis PCR-ribotyping protocol for *Clostridium difficile*", *PLoS ONE*, Jg. 10, Nr. 2, Feb. 2015. DOI: 10.1371/journal.pone.0118150.

[60] S. L. Stubbs, J. S. Brazier, G. L. O'Neill und B. I. Duerden, "PCR targeted to the 16S-23S rRNA gene intergenic spacer region of Clostridium difficile¡/i¿ and construction of a library consisting of 116 different PCR ribotypes", *Journal of Clinical Microbiology*, Nr. 2, S. 461–463, DOI: 10.1128/jcm.37.2.461-463.1999.

[61] M. C. J. Maiden, M. J. Jansen van Rensburg, J. E. Bray, S. G. Earle, S. a. Ford, K. a. Jolley und N. D. McCarthy, "MLST revisited: the gene-by-gene approach to bacterial genomics", *Nature reviews. Microbiology*, Jg. 11, Nr. 10, S. 728–36, 2013. DOI: 10.1038/nrmicro3093.

[62] K. A. Jolley und M. C. Maiden, "BIGSdb: Scalable analysis of bacterial genome variation at the population level", *BMC Bioinformatics*, Jg. 11, S. 595, Dez. 2010. DOI: 10.1186/1471-2105-11-595.

[63] K. A. Jolley, C. M. Bliss, J. S. Bennett, H. B. Bratcher, C. Brehony, F. M. Colles, H. Wimalarathna, O. B. Harrison, S. K. Sheppard, A. J. Cody und M. C. J. Maiden, "Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain", *Microbiology (Reading, England)*, Jg. 158, S. 1005–15, Apr. 2012. DOI: 10.1099/mic.0.055459-0.

[64] D. W. Eyre, T. E. Peto, D. W. Crook, A. Sarah Walker und M. H. Wilcox, "Hash-based core genome multilocus sequence typing for *Clostridium difficile*", *Journal of Clinical Microbiology*, Jg. 58, Nr. 1, Okt. 2020. DOI: 10.1128/JCM.01037-19.

[65] M. Frentrup, Z. Zhou, M. Steglich, J. P. Meier-Kolthoff, M. Göker, T. Riedel, B. Bunk, C. Spröer, J. Overmann, M. Blaschitz, A. Indra, L. von Müller, T. A. Kohl, S. Niemann, C. Seyboldt, F. Klawonn, N. Kumar, T. D. Lawley, S. García-Fernández, R. Cantón, R. del Campo, O. Zimmermann, U. Groß, M. Achtman und U. Nübel, "A publicly accessible database for *Clostridioides difficile* genome sequences supports tracing of transmission chains and epidemics", *In revision*, 2020.

[66] BioNumerics®, "*Clostridium difficile* schema for whole genome typing", *Release Note*,

[67] S. Bletz, S. Janezic, D. Harmsen, M. Rupnik und A. Mellmann, "Defining and evaluating a core genome multilocus sequence typing scheme for genome-wide typing of *Clostridium difficile*", *Journal of Clinical Microbiology*, Jg. 56, Nr. 6, e01987–17, Juni 2018. DOI: 10.1128/JCM.01987-17.

[68] Z. Zhou, N. F. Alikhan, K. Mohamed, Y. Fan und M. Achtman, "The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity", *Genome research*, Jg. 30, Nr. 1, S. 138–152, 2020. DOI: 10.1101/gr.251678.119.

[69] M. E. Pearce, N. F. Alikhan, T. J. Dallman, Z. Zhou, K. Grant und M. C. Maiden, "Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar enteritidis outbreak", *International Journal of Food Microbiology*, Jg. 274, S. 1–11, Juni 2018. DOI: 10.1016/j.ijfoodmicro.2018.02.023.

[70] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton und M. C. Frith, "Adaptive seeds tame genomic sequence comparison", *Genome Research*, Jg. 21, Nr. 3, S. 487–493, März 2011. DOI: 10.1101/gr.113985.110.

[71] X. Didelot, D. W. Eyre, M. Cule, C. L. C. Ip, M. A. Ansari, D. Griffiths, A. Vaughan, L. O'Connor, T. Golubchik, E. M. Batty, P. Piazza, D. J. Wilson, R. Bowden, P. J. Donnelly, K. E. Dingle, M. Wilcox, A. S. Walker, D. W. Crook, T. E. A. Peto und R. M. Harding, "Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission", *Genome biology*, Jg. 13, Nr. 12, R118, Dez. 2012. DOI: 10.1186/gb-2012-13-12-r118.

[72] D. W. Eyre, M. L. Cule, D. J. Wilson, D. Griffiths, A. Vaughan, L. O'Connor, C. L. Ip, T. Golubchik, E. M. Batty, J. M. Finney, D. H. Wyllie, X. Didelot, P. Piazza, R. Bowden, K. E. Dingle, R. M. Harding, D. W. Crook, M. H. Wilcox, T. E. Peto und A. S. Walker, "Diverse sources of *C. difficile* infection identified on whole-genome sequencing", *New England Journal of Medicine*, Jg. 369, Nr. 13, S. 1195–1205, 2013. DOI: 10.1056/NEJMoa1216064.

[73] D. R. Knight, M. M. Squire, D. A. Collins und T. V. Riley, "Genome analysis of *Clostridium difficile* PCR ribotype 014 lineage in Australian pigs and humans reveals a diverse genetic repertoire and signatures of long-range interspecies transmission", *Frontiers in microbiology*, Jg. 7, S. 2138, 2017. DOI: 10.3389/fmicb.2016.02138.

[74] X. Didelot und D. J. Wilson, "ClonalFrameML: Efficient Inference of recombination in whole bacterial genomes", *PLoS computational biology*, Jg. 11, Nr. 2, e1004041, Feb. 2015. DOI: 10.1371/journal.pcbi.1004041.

[75] A. Altmann, P. Weber, D. Bader, M. Preuß, E. B. Binder und B. Müller-Myhsok, "A beginners guide to SNP calling from high-Throughput DNA-sequencing data", *Human Genetics*, Jg. 131, Nr. 10, S. 1541–1554, 2012. DOI: 10.1007/s00439-012-1213-z.

[76] S. J. Bush, D. Foster, D. W. Eyre, E. L. Clark, N. De Maio, L. P. Shaw, N. Stoesser, T. E. A. Peto, D. W. Crook und A. S. Walker, "Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines", *GigaScience*, Jg. 9, S. 1–21, 2020. DOI: 10.1093/gigascience/giaa007.

[77] A. R. Wattam, J. J. Davis, R. Assaf, S. Boisvert, T. Brettin, C. Bun, N. Conrad, E. M. Dietrich, T. Disz, J. L. Gabbard, S. Gerdes, C. S. Henry, R. W. Kenyon, D. Machi, C. Mao, E. K. Nordberg, G. J. Olsen, D. E. Murphy-Olson, R. Olson, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, V. Vonstein, A. Warren, F. Xia, H. Yoo und R. L. Stevens, "Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center", *Nucleic Acids Research*, Jg. 45, Nr. D1, S. D535–D542, 2017. DOI: 10.1093/nar/gkw1017.

[78] O. Schwengers, A. Hoek, M. Fritzenwanker, L. Falgenhauer, T. Hain, T. Chakraborty und A. Goesmann, "ASA^{3}P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates", *PLOS Computational Biology*, Jg. 16, Nr. 3, M. Pertea, Hrsg., e1007134, März 2020. DOI: 10.1371/journal.pcbi.1007134.

[79] M. Steglich, "Evolutionsgenetik und Phylogeographie von *Clostridium difficile* Ribotyp 027 in Deutschland", *Dissertation, Technische Universität Carolo-Wilhelmina zu Braunschweig*, 2017.

[80] T. V. Riley, S. Thean, G. Hool und C. L. Golledge, "First Australian isolation of epidemic *Clostridium difficile* PCR ribotype 027", *Medical Journal of Australia*, Jg. 190, Nr. 12, S. 706–708, Juni 2009. DOI: 10.5694/j.1326-5377.2009.tb02644.x.

[81] M. P. Bauer, D. W. Notermans, B. H. Van Benthem, J. S. Brazier, M. H. Wilcox, M. Rupnik, D. L. Monnet, J. T. Van Dissel und E. J. Kuijper, "*Clostridium difficile* infection in Europe: A hospital-based survey", *The Lancet*, Jg. 377, Nr. 9759, S. 63–73, Jan. 2011. DOI: 10.1016/S0140-6736(10)61266-4.

[82] P. Moono, N. F. Foster, D. J. Hampson, D. R. Knight, L. E. Bloomfield und T. V. Riley, "*Clostridium difficile* Infection in production animals and avian species: A review", *Foodborne Pathogens and Disease*, Jg. 13, Nr. 12, S. 647–655, Dez. 2016. DOI: 10.1089/fpd.2016.2181.

[83] D. R. Knight, P. Putsathit, B. Elliott und T. V. Riley, "Contamination of Australian newborn calf carcasses at slaughter with *Clostridium difficile*", *Clinical Microbiology and Infection*, Jg. 22, Nr. 3, 266.e1–266.e7, März 2016. DOI: 10.1016/j.cmi.2015.11.017.

[84] B. M. Lund und M. W. Peck, "A possible route for foodborne transmission of *Clostridium difficile*?", *Foodborne Pathogens and Disease*, Jg. 12, Nr. 3, S. 177–182, März 2015. DOI: 10.1089/fpd.2014.1842.

[85] C. Candel-Pérez, G. Ros-Berruezo und C. Martínez-Graciá, "A review of *Clostridioides [Clostridium] difficile* occurrence through the food chain", *Food Microbiology*, Jg. 77, S. 118–129, Feb. 2019. DOI: 10.1016/j.fm.2018.08.012.

[86] A. Schneeberg, H. Neubauer, G. Schmoock, S. Baier, J. Harlizius, H. Nienhoff, K. Brase, S. Zimmermann und C. Seyboldt, "*Clostridium difficile* genotypes in piglet populations in Germany", *Journal of Clinical Microbiology*, Jg. 51, Nr. 11, S. 3796–3803, 2013. DOI: 10.1128/JCM.01440-13.

[87] C. W. Knetsch, N. Kumar, S. C. Forster, T. R. Connor, H. P. Browne, C. Harmanus, I. M. Sanders, S. R. Harris, L. Turner, T. Morris, M. Perry, F. Miyajima, P. Roberts, M. Pirmohamed, J. G. Songer, J. S. Weese, A. Indra, J. Corver, M. Rupnik, B. W. Wren, T. V. Riley, E. J. Kuijper und T. D. Lawley, "Zoonotic transfer of *Clostridium difficile* harboring antimicrobial resistance between farm animals and humans", *Journal of Clinical Microbiology*, Jg. 56, Nr. 3, März 2018. DOI: 10.1128/JCM.01384-17.

[88] V. Zidaric, M. Zemljic, S. Janezic, A. Kocuvan und M. Rupnik, "High diversity of *Clostridium difficile* genotypes isolated from a single poultry farm producing replacement laying hens", *Anaerobe*, Jg. 14, Nr. 6, S. 325–327, Dez. 2008. DOI: 10.1016/j.anaerobe.2008.10.001.

[89] C. Simango, "Prevalence of *Clostridium difficile* in the environment in a rural community in Zimbabwe", *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Jg. 100, Nr. 12, S. 1146–1150, Dez. 2006. DOI: 10.1016/j.trstmh.2006.01.009.

[90] D. Numberger, T. Riedel, G. McEwen, U. Nübel, M. Frentrup, I. Schober, B. Bunk, C. Spröer, J. Overmann, H. P. Grossart und A. D. Greenwood, "Genomic analysis of three *Clostridioides difficile* isolates from urban water sources", *Anaerobe*, Jg. 56, S. 22–26, 2019. DOI: 10.1016/j.anaerobe.2019.01.002.

[91] V. Pasquale, V. J. Romano, M. Rupnik, S. Dumontet, I. Čižnár, F. Aliberti, F. Mauri, V. Saggiomo und K. Krovacek, "Isolation and characterization of *Clostridium difficile* from shellfish and marine environments", *Folia Microbiologica*, Jg. 56, Nr. 5, S. 431–437, Sep. 2011. DOI: 10.1007/s12223-011-0068-3.

[92] R. B. Harvey, K. N. Norman, K. Andrews, M. E. Hume, C. M. Scanlan, T. R. Callaway, R. C. Anderson und D. J. Nisbet, "*Clostridium difficile* in poultry and poultry meat", *Foodborne Pathogens and Disease*, Jg. 8, Nr. 12, S. 1321–1323, Dez. 2011. DOI: 10.1089/fpd.2011.0936.

[93] S. P. Borriello, P. Honour, T. Turner und F. Barclay, "Household pets as a potential reservoir for *Clostridium difficile* infection", *Journal of Clinical Pathology*, Jg. 36, Nr. 1, S. 84–87, Jan. 1983. DOI: 10.1136/jcp.36.1.84.

[94] S. Janezic, S. Mlakar und M. Rupnik, "Dissemination of *Clostridium difficile* spores between environment and households: Dog paws and shoes", *Zoonoses and Public Health*, Jg. 65, Nr. 6, S. 669–674, Sep. 2018. DOI: 10.1111/zph.12475.

[95] M. Usui, M. Kawakura, N. Yoshizawa, L. L. San, C. Nakajima, Y. Suzuki und Y. Tamura, "Survival and prevalence of *Clostridium difficile* in manure compost derived from pigs", *Anaerobe*, Jg. 43, S. 15–20, Feb. 2017. DOI: 10.1016/j.anaerobe.2016.11.004.

[96] M. Dharmasena und X. Jiang, "Isolation of toxigenic *Clostridium difficile* from animal manure and composts being used as biological soil amendments", *Applied and Environmental Microbiology*, Jg. 84, Nr. 16, Juni 2018. DOI: 10.1128/aem.00738-18.

[97] D. Griffiths, W. Fawley, M. Kachrimanidou, R. Bowden, D. W. Crook, R. Fung, T. Golubchik, R. M. Harding, K. J. M. Jeffery, K. A. Jolley, R. Kirton, T. E. Peto, G. Rees, N. Stoesser, A. Vaughan, A. S. Walker, B. C. Young, M. Wilcox und K. E. Dingle, "Multilocus sequence typing of *Clostridium difficile*", *Journal of Clinical Microbiology*, Jg. 48, Nr. 3, S. 770–778, März 2010. DOI: 10.1128/JCM.01796-09.

[98] K. E. Dingle, D. Griffiths, X. Didelot, J. Evans, A. Vaughan, M. Kachrimanidou, N. Stoesser, K. A. Jolley, T. Golubchik, R. M. Harding, T. E. Peto, W. Fawley, A. S. Walker, M. Wilcox und D. W. Crook, "Clinical *Clostridium difficile*: clonality and pathogenicity locus diversity", *PLoS ONE*, Jg. 6, Nr. 5, O. Neyrolles, Hrsg., e19993, Mai 2011. DOI: 10.1371/journal.pone.0019993.

[99] D. W. Eyre, L. Tracey, B. Elliott, C. Slimings, P. G. Huntington, R. L. Stuart, T. M. Korman, G. Kotsiou, R. McCann, D. Griffiths, W. N. Fawley, P. Armstrong, K. E. Dingle, A. S. Walker, T. E. Peto, D. W. Crook, M. H. Wilcox und T. V. Riley, "Emergence and spread of predominantly community-onset *Clostridium difficile* PCR ribotype 244 infection in Australia, 2010 to 2012", *Euro Surveill*, Jg. 20, Nr. 10, S. 21 059, 2015.

[100] S. Polivkova, M. Krutova, K. Petrlova, J. Benes und O. Nyc, "*Clostridium difficile* ribotype 176 - A predictor for high mortality and risk of nosocomial spread?", *Anaerobe*, Jg. 40, S. 35–40, Aug. 2016. DOI: 10.1016/j.anaerobe.2016.05.002.

[101] K. Imwattana, D. R. Knight, B. Kullin, D. A. Collins, P. Putsathit, P. Kiratisin und T. V. Riley, "*Clostridium difficile* ribotype 017 – characterization, evolution and epidemiology of the dominant strain in Asia", *Emerging Microbes & Infections*, Jg. 8, Nr. 1, S. 796–807, 2019. DOI: 10.1080/22221751.2019.1621670.

[102] D. A. Collins, P. M. Hawkey und T. V. Riley, "Epidemiology of *Clostridium difficile* infection in Asia", *Antimicrobial Resistance and Infection Control*, Jg. 2, Nr. 1, S. 21, Juli 2013. DOI: 10.1186/2047-2994-2-21.

[103] M. D. Cairns, M. D. Preston, C. L. Hall, D. N. Gerding, P. M. Hawkey, H. Kato, H. Kim, E. J. Kuijper, T. D. Lawley, H. Pituch, S. Reid, B. Kullin, T. V. Riley, K. Solomon, P. J. Tsai, J. S. Weese, R. A. Stabler und B. W. Wren, "Comparative genome analysis and global phylogeny of the toxin variant *Clostridium difficile* PCR ribotype 017 reveals the evolution of two independent sublineages", *Journal of Clinical Microbiology*, Jg. 55, Nr. 3, S. 865–876, 2017. DOI: 10.1128/JCM.01296-16.

[104] I. A. Tickler, R. V. Goering, J. D. Whitmore, A. N. Lynn, D. H. Persing und F. C. Tenover, "Strain types and antimicrobial resistance patterns of *Clostridium difficile* isolates from the United States, 2011 to 2013", *Antimicrobial Agents and Chemotherapy*, Jg. 58, Nr. 7, S. 4214–4218, 2014. DOI: 10.1128/AAC.02775-13.

[105] A. C. Cheng, D. A. Collins, B. Elliott, J. K. Ferguson, D. L. Paterson, S. Thean und T. V. Riley, "Laboratory-based surveillance of *Clostridium difficile* circulating in Australia, September - November 2010", *Pathology*, Jg. 48, Nr. 3, S. 257–260, Apr. 2016. DOI: 10.1016/j.pathol.2016.02.005.

[106] R. Koch-institut, "Ausbruchsuntersuchungen bei *Clostridium (Clostridioides) difficile*", *Epidemiologisches Bulletin*, Jg. 14, S. 138–142, 2018. DOI: 10.1016/j.cmi.2017.10.008.

[107] J. M. Besser, H. A. Carleton, E. Trees, S. G. Stroika, K. Hise, M. Wise und P. Gerner-Smidt, "Interpretation of whole-genome sequencing for enteric disease surveillance and outbreak investigation", *Foodborne Pathogens and Disease*, Jg. 16, Nr. 7, S. 504–512, Juli 2019. DOI: 10.1089/fpd.2019.2650.

[108] J. Stimson, J. Gardy, B. Mathema, V. Crudu, T. Cohen und C. Colijn, "Beyond the SNP threshold: Identifying outbreak clusters using inferred transmissions", *Molecular Biology and Evolution*, Jg. 36, Nr. 3, S. 587–603, 2019. DOI: 10.1093/molbev/msy242.

[109] T. J. Dallman, P. M. Ashton, L. Byrne, N. T. Perry, L. Petrovska, R. Ellis, L. Allison, M. Hanson, A. Holmes, G. J. Gunn, M. E. Chase-Topping, M. E. J. Woolhouse, K. A. Grant, D. L. Gally, J. Wain und C. Jenkins, "Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK", *Microbial Genomics*, Jg. 1, Nr. 3, Sep. 2015. DOI: 10.1099/mgen.0.000029.

[110] S. Octavia, Q. Wang, M. M. Tanaka, S. Kaur, V. Sintchenko und R. Lan, "Delineating community outbreaks of *Salmonella enterica* serovar typhimurium by use of whole-genome sequencing: Insights into genomic variability within an outbreak", *Journal of Clinical Microbiology*, Jg. 53, Nr. 4, S. 1063–1071, Apr. 2015. DOI: 10.1128/JCM.03235-14.

[111] A. X. Han, E. Parker, S. Maurer-Stroh und C. A. Russell, "Inferring putative transmission clusters with Phydelity", *Virus Evolution*, Jg. 5, Nr. 2, Juli 2019. DOI: 10.1093/ve/vez039.

[112] H. Jia, P. Du, H. Yang, Y. Zhang, J. Wang, W. Zhang, G. Han, N. Han, Z. Yao, H. Wang, J. Zhang, Z. Wang, Q. Ding, Y. Qiang, F. Barbut, G. F. Gao, Y. Cao, Y. Cheng und C. Chen, "Nosocomial transmission of *Clostridium difficile* ribotype 027 in a Chinese hospital, 2012-2014, traced by whole genome sequencing", *BMC Genomics*, Jg. 17, Nr. 1, S. 405, 2016. DOI: 10.1186/s12864-016-2708-0.

[113] J. Revez, L. Espinosa, B. Albiger, K. C. Leitmeyer und M. J. Struelens, "Survey on the use of whole-genome sequencing for infectious diseases surveillance: Rapid expansion of European national capacities, 2015–2016", *Frontiers in Public Health*, Jg. 5, S. 347, Dez. 2017. DOI: 10.3389/fpubh.2017.00347.

[114] N. F. Alikhan, Z. Zhou, M. J. Sergeant und M. Achtman, "A genomic overview of the population structure of *Salmonella*", *PLoS Genetics*, Jg. 14, Nr. 4, e1007261, 2018. DOI: 10.1371/journal.pgen.1007261.

[115] Z. Zhou, N. F. Alikhan, M. J. Sergeant, N. Luhmann, C. Vaz, A. P. Francisco, J. A. Carriço und M. Achtman, "Grapetree: Visualization of core genomic relationships among 100,000 bacterial pathogens", *Genome Research*, Jg. 28, Nr. 9, S. 1395–1404, 2018. DOI: 10.1101/gr.232397.117.

[116] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. Adresse: https://www.r-project.org/.

[117] N. Thiel, S. Münch, W. Behrens, V. Junker, M. Faust, O. Biniasch, T. Kabelitz, P. Siller, C. Boedeker, P. Schumann, U. Rösler, T. Amon, K. Schepanski, R. Funk und U. Nübel, "Airborne bacterial emission fluxes from manure-fertilized agricultural soil", *submitted*, 2020.

[118] M. Baym, S. Kryazhimskiy, T. D. Lieberman, H. Chung, M. M. Desai, R. Kishony, G. Tyson, J. Chapman, P. Hugenholtz, E. Allen, R. Ram, M. Hegreness, R. Kishony, H. Bik, D. Porazinska, S. Creer, J. Caporaso, R. Knight, J. Barrick, R. Lenski, S. Kryazhimskiy, D. Rice, E. Jerison, M. Desai, P. Bielecki, P. Lukat, K. Hüsecken, A. Dötsch, H. Steinmetz, P. McAdam, E. Richardson, J. Fitzgerald, X. Didelot, R. Bowden, D. Wilson, T. Peto, D. Crook, V. Kuleshov, D. Xie, R. Chen, D. Pushkarev, Z. Ma, N. Rohland, D. Reich, S. Lamble, E. Batty, M. Attar, D. Buck, R. Bowden, A. Adey, H. Morrison, H. Morrison, X. Xun, J. Kitzman, A. Adey, J. Shendure, J. Thompson, L. Marcelino, M. Polz, M. DeAngelis, D. Wang und T. Hawkins, "Inexpensive multiplexed library preparation for megabase-sized genomes", *PLOS ONE*, Jg. 10, Nr. 5, S. J. Green, Hrsg., e0128036, Mai 2015. DOI: 10.1371/journal.pone.0128036.

[119] N. H. Zaiß, W. Witte und U. Nübel, "Fluoroquinolone resistance and *Clostridium difficile*, Germany", *Emerging Infectious Diseases*, Jg. 16, Nr. 4, S. 675–677, Apr. 2010. DOI: 10.3201/eid1604.090859.

[120] F. K. Berger, S. Gfrörer, S. L. Becker, R. Baldan, D. M. Cirillo, M. Frentrup, M. Steglich, P. Engling, U. Nübel, A. Mellmann, M. Bischoff, B. Gärtner und L. von Müller, "Hospital outbreak due to *Clostridium difficile* ribotype 018 (RT018) in Southern Germany", *International Journal of Medical Microbiology*, Jg. 309, Nr. 3-4, S. 189–193, 2019. DOI: 10.1016/j.ijmm.2019.03.001.

[121] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev und P. A. Pevzner, "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing", *Journal of Computational Biology*, Jg. 19, Nr. 5, S. 455–477, Mai 2012. DOI: 10.1089/cmb.2012.0021.

[122] H. Li und R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform", *Bioinformatics*, Jg. 26, Nr. 5, S. 589–595, Jan. 2010. DOI: 10.1093/bioinformatics/btp698.

[123] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young und A. M. Earl, "Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement", *PLoS ONE*, Jg. 9, Nr. 11, J. Wang, Hrsg., e112963, Nov. 2014. DOI: 10.1371/journal.pone.0112963.

[124] D. E. Wood und S. L. Salzberg, "Kraken: Ultrafast metagenomic sequence classification using exact alignments", *Genome Biology*, Jg. 15, Nr. 3, R46, März 2014. DOI: 10.1186/gb-2014-15-3-r46.

[125] S. F. Altschul, W. Gish, W. Miller, E. W. Myers und D. J. Lipman, "Basic local alignment search tool", *Journal of Molecular Biology*, Jg. 215, Nr. 3, S. 403–410, Okt. 1990. DOI: 10.1016/S0022-2836(05)80360-2.

[126] M. Frentrup, "distmatrix_to_distlist", *GitHub repository*, 2020. Adresse: https://github.com/Martinique-F/distmatrix_to_distlist.

[127] ——, "alleles_to_binary", *GitHub repository*, 2020. Adresse: https://github.com/Martinique-F/alleles_to_binary.

[128] Z. Zhou, "EToKi (Enterobase Tool Kit)", *GitHub repository*, 2019. Adresse: https://github.com/zheminzhou/EToKi/tree/1fef474ea24241eaf9d3dd0c75d4a1526a46fefc.

[129] T. Schlüter, "Enterodump", *GitHub repository*, 2019. Adresse: https://github.com/timoschlueter/enterodump.

[130] A. Stamatakis, "RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies", en, *Bioinformatics (Oxford, England)*, Jg. 30, Nr. 9, S. 1312–1313, Mai 2014. DOI: 10.1093/bioinformatics/btu033.

[131] Z. Zhou, A. McCann, F. X. Weill, C. Blin, S. Nair, J. Wain, G. Dougane und M. Achtman, "Transient darwinian selection in *Salmonella enterica* serovar paratyphi a during 450 years of global spread of enteric fever", *Proceedings of the National Academy of Sciences of the United States of America*, Jg. 111, Nr. 33, S. 12 199–12 204, Aug. 2014. DOI: 10.1073/pnas.1411012111.

[132] T. Seemann, F. Klötzl und A. J. Page, "snp-dists", *GitHub repository*, 2019. Adresse: https://github.com/tseemann/snp-dists.

[133] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM", *arXiv*, März 2013. arXiv: 1303.3997.

[134] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis und R. Durbin, "The sequence alignment/map format and SAMtools", *Bioinformatics*, Jg. 25, Nr. 16, S. 2078–2079, Aug. 2009. DOI: 10.1093/bioinformatics/btp352.

[135] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding und R. K. Wilson, "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing", *Genome research*, Jg. 22, Nr. 3, S. 568–76, März 2012. DOI: 10.1101/gr.129684.111.

[136] M. Frentrup, "pairwise-snps-list", *GitHub repository*, 2020. Adresse: https://github.com/Martinique-F/pairwise-snps-list.

[137] R. G. H. Pagès, P. Aboyoun und S. DebRoy, *Biostrings: Efficient manipulation of biological strings*, 2019. DOI: 10.18129/B9.bioc.Biostrings.

[138] C. W. Knetsch, T. R. Connor, A. Mutreja, S. M. van Dorp, I. M. Sanders, H. P. Browne, D. Harris, L. Lipman, E. C. Keessen, J. Corver, E. J. Kuijper und T. D. Lawley, "Whole genome sequencing reveals potential spread of *Clostridium difficile* between humans and farm animals in the Netherlands, 2002 to 2011", *Eurosurveillance*, Jg. 19, Nr. 45, S. 1–12, Nov. 2014. DOI: 10.2807/1560-7917.es2014.19.45.20954.

[139] R. A. Stabler, M. He, L. Dawson, M. Martin, E. Valiente, C. Corton, T. D. Lawley, M. Sebaihia, M. A. Quail, G. Rose, D. N. Gerding, M. Gibert, M. R. Popoff, J. Parkhill, G. Dougan und B. W. Wren, "Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium", *Genome Biology*, Jg. 10, Nr. 9, R102, Sep. 2009. DOI: 10.1186/gb-2009-10-9-r102.

[140] M. He, M. Sebaihia, T. D. Lawley, R. A. Stabler, L. F. Dawson, M. J. Martin, K. E. Holt, H. M. Seth-Smith, M. A. Quail, R. Rance, K. Brooks, C. Churcher, D. Harris, S. D. Bentley, C. Burrows, L. Clark, C. Corton, V. Murray, G. Rose, S. Thurston, A. Van Tonder, D. Walker, B. W. Wren, G. Dougan und J. Parkhill, "Evolutionary dynamics of *Clostridium difficile* over short and long time scales", *Proceedings of the National Academy of Sciences of the United States of America*, Jg. 107, Nr. 16, S. 7527–7532, 2010. DOI: 10.1073/pnas.0914322107.

[141] J. A. Lees, S. R. Harris, G. Tonkin-Hill, R. A. Gladstone, S. W. Lo, J. N. Weiser, J. Corander, S. D. Bentley und N. J. Croucher, "Fast and flexible bacterial genomic epidemiology with PopPUNK", *Genome Research*, Jg. 29, Nr. 2, S. 304–316, Jan. 2019. DOI: 10.1101/gr.241455.118.

[142] N. Saitou, *Introduction to evolutionary genomics*, Ser. Computational Biology. Cham: Springer International Publishing, 2018, Bd. 17. DOI: 10.1007/978-3-319-92642-1.

[143] M. Simonsen, T. Mailund und C. N. S. Pedersen, "Rapid Neighbour-Joining", Techn. Ber., 2008. DOI: 0.1007/978-3-540-87361-7_10.

[144] J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs und H. Wagner, *vegan: Community Ecology Package*, 2019. Adresse: https://cran.r-project.org/package=vegan.

[145] J. A. Carriço, C. Silva-Costa, J. Melo-Cristino, F. R. Pinto, H. De Lencastre, J. S. Almeida und M. Ramirez, "Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*", *Journal of Clinical Microbiology*, Jg. 44, Nr. 7, S. 2524–2532, Juli 2006. DOI: 10.1128/JCM.02536-05.

[146] T. C. Hsieh, K. H. Ma und A. Chao, "iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers)", *Methods in Ecology and Evolution*, Jg. 7, Nr. 12, S. 1451–1456, 2016. DOI: 10.1111/2041-210X.12613.

[147] M. J. Crawley, *The R Book*. Chichester, UK: John Wiley und Sons, Sep. 2007, S. 1–942. DOI: 10.1002/9780470515075.

[148] J. P. Meier-Kolthoff, H. P. Klenk und M. Göker, "Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age", *International Journal of Systematic and Evolutionary Microbiology*, Jg. 64, Nr. PART 2, S. 352–356, 2014. DOI: 10.1099/ijs.0.056994-0.

[149] J. Hausser und K. Strimmer, *entropy: Estimation of entropy, mutual information and related quantities*, 2014. Adresse: https://rdrr.io/cran/entropy/.

[150] H. Wickham, R. François, L. Henry und K. Müller, *dplyr: A grammar of data manipulation*, 2020. Adresse: http://dplyr.tidyverse.org.

[151] E. Herberich, J. Sikorski und T. Hothorn, "A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs", *PLoS ONE*, Jg. 5, Nr. 3, S. 1–8, 2010. DOI: 10.1371/journal.pone.0009788.

[152] K. E. Dingle, B. Elliott, E. Robinson, D. Griffiths, D. W. Eyre, N. Stoesser, A. Vaughan, T. Golubchik, W. N. Fawley, M. H. Wilcox, T. E. Peto, A. S. Walker, T. V. Riley, D. W. Crook und X. Didelot, "Evolutionary history of the *Clostridium difficile* pathogenicity locus", *Genome biology and evolution*, Jg. 6, Nr. 1, S. 36–52, Jan. 2014. DOI: 10.1093/gbe/evt204.

[153] D. W. Eyre, W. N. Fawley, A. Rajgopal, C. Settle, K. Mortimer, S. D. Goldenberg, S. Dawson, D. W. Crook, T. E. A. Peto, A. S. Walker und M. H. Wilcox, "Comparison of control of *Clostridium difficile* infection in six English hospitals using whole-genome sequencing", *Clinical Infectious Diseases*, Jg. 65, Nr. 3, S. 433–441, 2017. DOI: 10.1093/cid/cix338.

[154] S. Janezic und M. Rupnik, "Development and implementation of whole genome sequencing-based typing schemes for *Clostridioides difficile*", *Frontiers in Public Health*, Jg. 7, Okt. 2019. DOI: 10.3389/fpubh.2019.00309.

[155] N. Kumar, F. Miyajima, M. He, P. Roberts, A. Swale, L. Ellison, D. Pickard, G. Smith, R. Molyneux, G. Dougan, J. Parkhill, B. W. Wren, C. M. Parry, M. Pirmohamed und T. D. Lawley, "Genome-based infection tracking reveals dynamics of *Clostridium difficile* transmission and disease recurrence", *Clinical Infectious Diseases*, Jg. 62, Nr. 6, S. 746–752, März 2015. DOI: 10.1093/cid/civ1031.

[156] K. E. Dingle, X. Didelot, T. P. Quan, D. W. Eyre, N. Stoesser, T. Golubchik, R. M. Harding, D. J. Wilson, D. Griffiths, A. Vaughan, J. M. Finney, D. H. Wyllie, S. J. Oakley, W. N. Fawley, J. Freeman, K. Morris, J. Martin, P. Howard, S. Gorbach, E. J. Goldstein, D. M. Citron, S. Hopkins, R. Hope, A. P. Johnson, M. H. Wilcox, T. E. Peto, A. S. Walker, D. W. Crook, C. Del Ojo Elias, C. Crichton, V. Kostiou, A. Giess und J. Davies, "Effects of control interventions on *Clostridium difficile* infection in England: An observational study", *The Lancet Infectious Diseases*, Jg. 17, Nr. 4, S. 411–421, Apr. 2017. DOI: 10.1016/S1473-3099(16)30514-X.

[157] D. W. Eyre, A. S. Walker, J. Freeman, S. D. Baines, W. N. Fawley, C. H. Chilton, D. Griffiths, A. Vaughan, D. W. Crook, T. E. A. Peto und M. H. Wilcox, "Short-term genome stability of serial *Clostridium difficile* ribotype 027 isolates in an experimental gut model and recurrent human disease", *PLoS ONE*, Jg. 8, Nr. 5, U. Nübel, Hrsg., e63540, Mai 2013. DOI: 10.1371/journal.pone.0063540.

[158] SeqSphere, *Tutorial for SeqSphere+ assembly and cgMLST analysis pipeline*. Adresse: https://www.ridom.de/seqsphere/ug/latest/Tutorial_pipeline.html.

[159] A. W. Pightling, N. Petronella und F. Pagotto, "Choice of reference-guided sequence assembler and SNP caller for analysis of *Listeria monocytogenes* short-read sequence data greatly influences rates of error genomics", *BMC Research Notes*, Jg. 8, Nr. 1, S. 748, Dez. 2015. DOI: 10.1186/s13104-015-1689-4.

[160] B. Jagadeesan, L. Baert, M. Wiedmann und R. H. Orsi, "Comparative analysis of tools and approaches for source tracking *Listeria monocytogenes* in a food facility using whole-genome sequence data", *Frontiers in Microbiology*, Jg. 10, Nr. MAY, S. 947, Mai 2019. DOI: 10.3389/fmicb.2019.00947.

[161] T. J. Treangen, B. D. Ondov, S. Koren und A. M. Phillippy, "The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes", *Genome biology*, Jg. 15, Nr. 11, S. 524, 2014. DOI: 10.1186/PREACCEPT-2573980311437212.

[162] A. C. Schürch, S. Arredondo-Alonso, R. J. Willems und R. V. Goering, "Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches", *Clinical Microbiology and Infection*, Jg. 24, Nr. 4, S. 350–354, Apr. 2018. DOI: 10.1016/j.cmi.2017.12.016.

[163] D. W. Eyre, "runListCompare", *GitHub repository*, 2019. Adresse: https://github.com/davideyre/runListCompare.

[164] R. S. Schwartz und R. L. Mueller, "Branch length estimation and divergence dating: Estimates of error in Bayesian and maximum likelihood frameworks", *BMC Evolutionary Biology*, Jg. 10, Nr. 1, S. 5, 2010. DOI: 10.1186/1471-2148-10-5.

[165] Z. Ruan, Y. Yu und Y. Feng, "The global dissemination of bacterial infections necessitates the study of reverse genomic epidemiology", *Briefings in Bioinformatics*, Jg. 00, Nr. November 2018, S. 1–10, 2019. DOI: 10.1093/bib/bbz010.

[166] K. M. Robinson, A. S. Hawkins, I. Santana-Cruz, R. S. Adkins, A. C. Shetty, S. Nagaraj, L. Sadzewicz, L. J. Tallon, D. A. Rasko, C. M. Fraser, A. Mahurkar, J. C. Silva und J. C. Hotopp, "Aligner optimization increases accuracy and decreases compute times in multi-species sequence data", *Microbial Genomics*, Jg. 3, Nr. 9, Sep. 2017. DOI: 10.1099/mgen.0.000122.

[167] D. W. Eyre, K. A. Davies, G. Davis, W. N. Fawley, K. E. Dingle, N. De Maio, A. Karas, D. W. Crook, T. E. A. Peto, A. S. Walker, M. H. Wilcox und E. S. EUCLID Study Group, "Two distinct patterns of *Clostridium difficile* diversity across Europe indicating contrasting routes of spread", *Clinical infectious diseases*, Jg. 67, Nr. 7, S. 1035–1044, 2018. DOI: 10.1093/cid/ciy252.

[168] L. C. McDonald, B. Coignard, E. Dubberke, X. Song, T. Horan und P. K. Kutty, "Recommendations for surveillance of *Clostridium difficile* –associated disease", *Infection Control Hospital Epidemiology*, Jg. 28, Nr. 2, S. 140–145, Feb. 2007. DOI: 10.1086/511798.

[169] L. C. McDonald, D. N. Gerding, S. Johnson, J. S. Bakken, K. C. Carroll, S. E. Coffin, E. R. Dubberke, K. W. Garey, C. V. Gould, C. Kelly, V. Loo, J. Shaklee Sammons, T. J. Sandora und M. H. Wilcox, *Clinical practice guidelines for Clostridium difficile infection in adults and children: 2017 update by the infectious diseases society of America (IDSA) and society for healthcare epidemiology of America (SHEA)*, 2018. DOI: 10.1093/cid/cix1085.

[170] Y. Longtin, B. Paquet-Bolduc, R. Gilca, C. Garenc, E. Fortin, J. Longtin, S. Trottier, P. Gervais, J. F. Roussy, S. Levesque, D. Ben-David, I. Cloutier und V. G. Loo, "Effect of detecting and isolating *Clostridium difficile* carriers at hospital admission on the incidence of C difficile infections; A quasi-experimental controlled study", *JAMA Internal Medicine*, Jg. 176, Nr. 6, S. 796–804, 2016. DOI: 10.1001/jamainternmed.2016.0177.

[171] D. W. Eyre, D. Griffiths, A. Vaughan, T. Golubchik, M. Acharya, L. O'Connor, D. W. Crook, A. S. Walker und T. E. Peto, "Asymptomatic *Clostridium difficile* colonisation and onward transmission", *PLoS ONE*, Jg. 8, Nr. 11, Y.-F. Chang, Hrsg., e78445, Nov. 2013. DOI: 10.1371/journal.pone.0078445.

[172] M. Krutova, P. Kinross, F. Barbut, A. Hajdu, M. H. Wilcox, E. J. Kuijper, F. Allerberger, M. Delmée, J. Van Broeck, R. Vatcheva-Dobrevska, E. Dobreva, B. Matica, D. Pieridou, M. Krůtová, O. Nyč, B. Olesen, P. Märtin, S. Mentula, F. Barbut, M. Arvand, L. von Müller, J. Papaparaskevas, J. Pászti, Hajdu, T. Gudnason, K. Burns, P. Spigaglia, K. Vulāne, M. Debacker, E. Scicluna, T. Melillo, E. J. Kuijper, M. T. Crobach, O. Kacelnik, E. Astrup, H. Pituch, M. Oleastro, C. Wiuff, J. Coia, E. Nováková, J. Kolman, E. Grilc, M. Rupnik, E. Bouza, E. Reigadas, T. Åkerlund, S. Tschudin-Sutter, M. H. Wilcox, D. Fairley und T. Morris, *How to: Surveillance of Clostridium difficile infections*, Mai 2018. DOI: 10.1016/j.cmi.2017.12.008.

[173] S. Janezic, M. Ocepek, V. Zidaric und M. Rupnik, "*Clostridium difficile* genotypes other than ribotype 078 that are prevalent among human, animal and environmental isolates", *BMC microbiology*, Jg. 12, Nr. 1, S. 48, März 2012. DOI: 10.1186/1471-2180-12-48.

[174] H. Kurka, A. Ehrenreich, W. Ludwig, M. Monot, M. Rupnik, F. Barbut, A. Indra, B. Dupuy und W. Liebl, "Sequence similarity of *Clostridium difficile* strains by analysis of conserved genes and genome content is reflected by their ribotype affiliation", *PLoS ONE*, Jg. 9, Nr. 1, M. M. Heimesaat, Hrsg., e86535, Jan. 2014. DOI: 10.1371/journal.pone.0086535.

Danksagung

An erster Stelle gilt mein großer Dank meinem Mentor Professor Dr. Ulrich Nübel für die Chance an solch einem abwechslungsreichen und interessantem Thema zu arbeiten, für die zahl- und lehrreichen Diskussionen, die gute Zusammenarbeit, die Möglichkeit die Arbeit auf Konferenzen vorzustellen und das ich in die Welt der Bioinformatik eintreten konnte.

Mein besonderer Dank gilt auch Professor Dr. Michael Steinert von der TU Braunschweig, zum einen für die Übernahme des Zweitgutachtens und zum anderen für die freundlichen Gespräche und konstruktiven Kommentare während des Promotionsbeirats.

Vielen Dank auch an Professor Dr. Miguel Vences von der TU Braunschweig für die Bereitschaft der Übernahme des Vorsitzes der Prüfungskommission.

Vielen Dank an Professor Dr. Jörg Overmann für die Möglichkeit die Arbeit am Leibniz-Institut DSMZ anfertigen zu können, die Bereitschaft am Promotionsbeirat teilzunehmen und die hilfreichen Anmerkungen.

Special thanks to Professor Dr. Mark Achtman, Dr. Zhemin Zhou, Dr. Nabil Farid Alikhan and Dr. Martin Sergeant for the development of EnteroBase, the quick and uncomplicated answers to my questions and for a very educational visit to the University of Warwick.

Besonders erwähnen möchte ich hier auch alle Kooperationspartner und Co-Autoren, die Isolate zur Verfügung gestellt und somit die Analysen dieser Arbeit erst ermöglichten. Danke an Professor Dr. med. Lutz von Müller, Dr. Christian Seyboldt, Professor Dr. med. Uwe Groß und besonders Ortrud Zimmermann, die oft auch sehr kurzfristig epidemiologische Daten zur Verfügung stellte. Danke an Dr. Alexander Indra und Marion Blaschitz für die Ribotypisierung unserer Isolate. Vielen Dank auch an Prof. Dr. Stefan Niemann und Dr. Thomas Kohl für die Sequenzierung einiger Isolate.
Prof. Dr. Frank Klawonn möchte ich für die sehr lehrreichen Diskussionen und die intensive Beratung zur statistischen Auswertung danken.

A big thank you to the cooperation partners at the University Clinic and the Ramón y Cajal University Hospital in Madrid, especially to Sergio García-Fernández for the really nice time in Madrid!

Dr. Thomas Riedel möchte ich dafür danken, dass er Ganzgenome von *C. difficile* für die Entwicklung der Datenbank und des cgMLST Schemas in EnteroBase zur Verfügung stellte.

Ein besonderer Dank gilt dem Betriebsrat der DSMZ und all seinen Mitgliedern für die große Unterstützung gerade am Ende dieser Arbeit, die vielen interessanten Diskussionen und die Möglichkeit mich für die Doktoranden an der DSMZ einsetzen zu können.

Danke auch an die Bioinformatik-Abteilung, besonders Dr. Boyke Bunk und Adam Podstawka für die oft sehr spontane Lösung von Serverproblemen und der großen Geduld bei der Beantwortung meiner Fragen.

Vielen Dank an Dr. Catrhin Spröer, Simone Severitt, Nicole Heyer und Carola Berg für die Sequenzierung der Genome und den vielen interessanten und hilfreichen Gesprächen.

Ich möchte mich auch bei allen Mitgliedern des R-Clubs an der DSMZ bedanken, besonders Dr. Johannes Sikorski für die Lösung vieler Probleme und interessanten und lehrreichen Diskussionen.

Ein ganz besonderer Dank geht an das gesamte Kollegium der DSMZ, für die tolle Arbeitsatmosphäre, die offenen Ohren und den netten Plausch auf dem Flur. Vielen Dank! Besonders hervorheben möchte ich hier:

- Dr. Matthias Steglich für seine unglaubliche Unterstützung in jeder Hinsicht, seine große Begeisterung für die Bioinformatik mit der er mich erfolgreich infiziert hat, den vielen geduldigen Erklärungen, die vielen interessanten Gespräche und dem guten Zureden in stressigen Situationen. Vielen Dank!

- Dr. Jan Meier-Kolthoff für die bioinformatische wie auch mentale Unterstützung und vielen anregenden Unterhaltungen.

- PD Dr. Markus Göker für die vielen geduldigen, ausführlichen Erklärungen und hilfreichen Skripte.

- Dr. Wiebke Behrens für die schöne, wenn auch kurze, Zusammenarbeit und hilfreichen Kommentaren für die Arbeit.

Mit dieser Arbeit schließe ich einen Lebensabschnitt ab, den ich ohne die Unterstützung meines inspirierenden Freundeskreises nicht so hätte zuende bringen können. Vielen lieben Dank an jeden einzelnen von euch und danke, dass ich euch zu meinen Freunden zählen darf! Besonders bedanken für die große Unterstützung während dieser Arbeit möchte ich mich bei:

- Dr. Felizitas Bajerski. Danke zunächst natürlich für das Korrekturlesen und die vielen hilfreichen Kommentare, sei es für diese Arbeit, für Vorträge oder in so ziemlich allen anderen Lebenslagen. Keine Ahnung wie ich die letzten 4 Jahre ohne dich überstanden hätte, du hast jedenfalls einen großen Anteil daran, dass ich die Zeit mit viel Freude und Spaß verbracht habe! Und natürlich auch danke an Randy für die intensiven Fußball- und Musikdiskussionen und die gute Verpflegung.

- Dr. Nadine Thiel. Danke für die vielen hilfreichen Anmerkungen bei dieser Arbeit, für die schöne gemeinsame Zeit an und abseits der DSMZ und den sehr spaßigen Konferenzbesuchen.

- Timo Schlüter. Danke für die informatische und mentale Unterstützung, fürs Korrekturlesen dieser Arbeit und das ständige Aufbauen meines Selbstbewusstseins.

- Vera Junker. Natürlich zunächst danke für die tolle Arbeit im Labor, wodurch ich überhaupt Sequenzen zum auswerten hatte. Vielen Dank aber auch für das allseits offene Ohr und die Möglichkeit für regelmäßige Felltherapie bei Sam.

- Carlo für durchtanzte Nächte und den 1-2 Bier zwischendurch, Amelie - geteiltes Leid ist halbes Leid, Steffen für entspannte Wochenenden, Anna und Carina für viele inspirierende Gespräche außerhalb des Forschungswahnsinns.

- Max und Katharina, die immer für mich da sind und mir in Duisburg immer ein Gefühl von Heimat und Entspannung geben. Danke für eure jahrelange Unterstützung!

Weiterhin möchte ich mich beim MSV Duisburg für die jahrelange und manchmal leider etwas zu intensive Ausbildung meines großen Frustrationspotentials bedanken, aber auch für die Lektion, dass man große Rückschläge und Stresssituationen mit viel Leidenschaft überstehen kann.

Die Unterstützung, die ich von meiner Familie über all die Jahre erfahren habe, lässt sich eigentlich nicht in Worte fassen. Danke an meine Tante Doris Terörde und meinen Onkel Josef Binder. Danke an meine Patentante Brigitte Frentrup, die mir mit ihrem Lebensweg vorgelebt hat, dass es sich lohnt auch mal neue Wege zu gehen. Besonders erwähnen möchte ich hier meinen Onkel Wilfried Binder, der mich mit seinem hilfsbereiten und allseits positiven Gemüt sehr geprägt hat und mir das Abenteuer Braunschweig überhaupt erst ermöglichte.
Meine Eltern, Klaus und Monika Frentrup, möchte ich hier nicht nur außerordentlich für all ihre Unterstützung über all die Jahre danken, sondern auch sagen, was für wunderbare Menschen sie sind. Danke, dass ich meinen eigenen Weg gehen durfte und ihr diesen mit mir gegangen seid.

Zu guter Letzt möchte ich mich bei mir selbst bedanken. Dafür, dass ich die meiste Zeit auf mich und meinen Körper geachtet und meine Leidenschaft für die Forschung auch während stressiger Phasen nicht verloren habe.